CFIEC AI Study Group Report:

Toward Establishing Certification and Legal Frameworks for Unsupervised and High-risk Activities AI (UHAI)

March of 2026

General Incorporated Foundation

Center for International Economic Collaboration (CFIEC)

CFIEC AI Study Group Report: Toward Establishing Evaluation, Certification, and Legal Frameworks for Unsupervised and High-risk Activities AI (UHAI)

Table of Contents

Executive Summary

1. The Potential of "UHAI" to Deliver Safety Beyond Human Capabilities

The capabilities of artificial intelligence (AI) are advancing at a pace that far exceeds our expectations. Since 2023, state-of-the-art AI systems have achieved scores well above passing thresholds on Japan's National Medical Licensing Examination and the bar examination (short-answer section), and in January 2026, an AI system achieved an overall score of 96.9% on the Common Test for University Admissions.

These capability gains extend far beyond examinations and dialogue. In medical imaging, AI has been demonstrated to substantially reduce miss rates for lung and breast cancer compared with multiple radiologists. In the automotive sector, autonomous driving systems have shown remarkable results, reducing serious injury crashes by 90% relative to human drivers.

In light of these realities, the conventional assumption that "safety is ensured as long as humans monitor and check AI" can no longer be taken for granted. Indeed, experimental evidence shows that in domains where AI outperforms humans, human involvement in AI decision-making can actually degrade performance. When humans "second-guess" a capable AI, outcomes worsen.

This report defines the concept of UHAI (Unsupervised and High-risk Activities AI)—AI that autonomously performs high-risk activities without direct human oversight—and sets out to describe the institutional architecture necessary for its social implementation. Typical examples include Level 4 and above autonomous vehicles and AI medical devices that perform diagnoses without physician involvement. The social implementation of UHAI holds the potential not merely to enhance efficiency but to deliver levels of safety unattainable by humans alone. Realizing this potential, however, requires overcoming three barriers related to the evaluation, certification, and legal frameworks for UHAI.

2. Three Institutional Barriers Impeding UHAI Implementation

UHAI possesses three technical characteristics: (i) a "black-box nature," whereby the decision-making process cannot be explained even by the developers themselves; (ii) "dynamic mutability," whereby system behavior continuously changes through learning; and (iii) "difficulty in goal-setting," whereby the thresholds for safety remain undefined. These characteristics give rise to the following challenges for existing institutional frameworks.

(i) Absence of evaluation criteria: Due to AI's black-box nature, conventional product inspection methods cannot demonstrate safety.

(ii) Inadequacy of certification systems: Even if evaluation criteria were established, the mechanisms and organizational capacity for independent third parties to “certify” AI safety are not yet in place.

(iii) Structural limitations of legal frameworks: Current regulatory regimes, enforcement mechanisms, and liability systems are all designed on the premise that “a human can ultimately make judgments and intervene.”

3. Current Institutional Achievements and Challenges (Case Studies)

This report distills three common directions from case studies of medical device AI (the IDATEN and PCCP frameworks in Japan and the United States) and autonomous driving (Japan's Specified Autonomous Driving system, the United Kingdom's Automated Vehicles Act, and California's deployment permit system). First, a two-stage evaluation that abandons efforts to elucidate AI's "internals" in favor of combining behavioral verification with organizational governance assessment. Second, a shift from one-time assessments to continuous management throughout the product lifecycle. Third, the construction of legal frameworks that treat the maintenance of appropriate management systems as a safe harbor (basis for immunity). However, each of these systems has unresolved challenges, and the development of evaluation, certification, and legal frameworks for UHAI remains a work in progress.

4. Recommendations of This Report

Based on the foregoing analysis, this report proposes the following three directions.

(1) Evaluation of UHAI

A new evaluation paradigm should be constructed that addresses the three technical characteristics of UHAI: black-box nature, dynamic mutability, and difficulty in goal-setting. With respect to the black-box nature, rather than attempting to elucidate AI's internals, techniques for externally verifying AI "behavior" should be standardized, leveraging new V&V methods such as formal methods and red teaming. With respect to dynamic mutability, pre-market verification and post-market real-world performance monitoring should be designed as an integrated system, establishing mechanisms for continuous evaluation throughout the entire lifecycle. With respect to the difficulty in goal-setting, rather than pursuing "absolute safety," acceptable risk levels should be formulated through a transparent process involving diverse stakeholders, anchored in relative and statistical comparison with conventional human operations, with that consensus serving as the starting point for evaluation criteria.

(2) Certification for UHAI

A certification system should be constructed to give effective backing to the above evaluation criteria. With respect to the black-box nature, a "joint certification" framework should be constructed that examines not only the product's input-output verification but also the governance structure of the organization that develops and operates it. With respect to dynamic mutability, rather than requiring a fresh review from scratch for each change, an approach that pre-approves the change management process itself should be advanced. With respect to the difficulty in goal-setting, socially agreed standards should be connected to certification judgment criteria, and the capacity for certification bodies to independently verify these should be built.

(3) Legal Frameworks for UHAI

It is important to legally accept the outcomes of evaluation and certification and to secure legal predictability for business operators. With respect to the black-box nature, the technical assurance of safe operating domains should be accepted as an authorization requirement in lieu of explanations of internal structure, and the maintenance of appropriate organizational structures should be positioned as a safe harbor. With respect to dynamic mutability, operators should be mandated to conduct continuous safety management and self-reporting, with the effectiveness of these obligations ensured through appropriate incentivization of information provision. With respect to the difficulty in goal-setting, test environments and certification systems should be developed that enable operators to demonstrate achievement of socially agreed standards.

These three layers must not be designed in isolation but must be designed and operated in an integrated manner. That is, consensus on evaluation criteria provides the foundation for certification; certification attainment provides the basis for safe harbors; and the clarification of safe harbors enhances operators' legal predictability and promotes innovation.

UHAI holds the potential to transcend the limits of human capability and deliver greater safety and benefits to society. To make this potential a reality, we must move beyond the conventional assumption that "safety is ensured by human oversight" and envision new institutional frameworks that can adapt to the evolution of technology. It is our hope that the issues organized in this report will serve as a starting point for that discussion.

Table: Technical Characteristics of UHAI (columns) and Corresponding Directions for Institutional Design (rows)

	Black-box nature	**Dynamic mutability**	**Difficulty in goal-setting**
Evaluation	Abandon elucidation of AI internals; **standardize techniques for externally verifying AI "behavior"** through new V&V methods such as formal methods and red teaming	**Design pre-market verification and post-market Real-World Performance (RWP) monitoring as an integrated system**, establishing continuous evaluation throughout the entire lifecycle	Rather than pursuing "absolute safety," build consensus on acceptable risk levels through a transparent process with diverse stakeholders, anchored in **relative and statistical comparison** with conventional human operations
Certification	Construct a **"joint certification" framework that reviews both** product input-output verification and organizational governance as a package	Rather than re-reviewing from scratch for each change, advance an approach that **pre-approves the change management process itself**	**Connect socially agreed standards to certification judgment criteria** and build the capacity for certification bodies to independently verify them
Legal frameworks	Accept **technical assurance of safe operating domains as an authorization requirement** in lieu of explaining internal structure; legally position maintenance of appropriate organizational structures as a **safe harbor**	**Mandate continuous safety management and self-reporting** by operators, ensuring effectiveness through appropriate incentivization of information provision	Develop **test environments and certification systems** that enable operators to demonstrate achievement of agreed standards, securing safe harbors that provide legal predictability

1. Background and Purpose of This Report: The Potential of "UHAI" to Deliver Safety Beyond Human Capabilities

1.1 The Remarkable Advances of Artificial Intelligence

We are living through an extraordinary moment in the history of technology.

In 1950, the father of computer science, Alan Turing, proposed the famous "Turing Test": if a human judge, communicating through text, could not distinguish between a human and a machine, the machine should be considered intelligent. For over seventy years, this test served both as the ultimate goal of AI research and as a symbol of an "insurmountable barrier." Yet in research since 2023, GPT-4 succeeded in being misidentified as human by a majority of judges, and by 2025, GPT-4.5 was judged to be human 73% of the time.[1]

AI has also begun recording scores far above passing thresholds on examinations designed to select highly qualified professionals. For example, in the January 2026 Common Test for University Admissions, the latest generation of AI achieved perfect scores in multiple core subjects and an overall score of 96.9%.[2] On the National Medical Licensing Examination, AI has consistently maintained accuracy rates comparable to practicing physicians, including on questions involving complex clinical reasoning, placing it firmly within the passing range.[3] Furthermore, on the 2025 bar examination (short-answer section), AI recorded an accuracy rate exceeding 96%, rivaling the top tier of human examinees.[4]

---

[1] Jones, C. & Bergen, B. (2024). People cannot distinguish GPT-4 from a human in a Turing test. arXiv:2405.08007. In this study, GPT-4 succeeded in being misidentified as human by 54% of judges. Furthermore, Jones, C. & Bergen, B. (2025). Large Language Models Pass the Turing Test. arXiv:2503.23674 reported that GPT-4.5 was judged to be human 73% of the time.

[2] Joint verification by LifePrompt, Inc. and Nikkei Inc. (published January 20, 2026). OpenAI's GPT-5.2 Thinking achieved a score rate of 96.9% across 15 subjects and perfect scores in 9 subjects on the 2026 Common Test for University Admissions. See Nihon Keizai Shimbun, January 21, 2026, "Common Test for University Admissions: OpenAI achieves perfect scores in 9 subjects, 97% score rate."

[3] Takagi, S., et al. (2023). Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Medical Education, 9, e48002; Nomura, A., et al. (2024). Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan. PLOS Digital Health, 3(1), e0000433. On the 119th National Medical Licensing Examination (2025), OpenAI o1 achieved a score rate of 98% on essential questions (Medic Media verification).

[4] Bengo4.com, Inc. "Legal Brain, a law-specialized AI, records 96.5% accuracy on the 2025 Bar Examination (Short-Answer Section)" (November 12, 2025). It scored 169 out of 175 points, exceeding the highest score among examinees (167 points).

The domains in which AI can outperform humans are not limited to dialogue and examination questions. Even in "high-risk decisions" that directly affect our lives and livelihoods—and that have accordingly been subject to strict regulation—AI is becoming a more accurate and reliable agent than humans. The following sections present specific examples in healthcare and autonomous driving.

1.1.1 Medical Imaging

Medical imaging is one of the domains where AI was earliest demonstrated to surpass human experts.

In lung cancer screening, a 2019 joint study by Google AI and Northwestern University reported that AI achieved diagnostic accuracy superior to all six experienced radiologists on low-dose CT scans. The AI reduced cancer misses (false negatives) by 5% and false detections (false positives) by 11%.[5]

Similar results have been reported in breast cancer screening. In 2020, a joint team from Google Health, DeepMind, and the UK National Health Service (NHS) published in *Nature* that AI reduced false positive rates by up to 5.7% and false negative rates by up to 9.4% in mammography. In a direct comparison with six radiologists, the AI outperformed all of them.[6]

In the United States, fully autonomous AI diagnostics have begun receiving regulatory approval. In 2018, the FDA approved the diabetic retinopathy diagnostic AI "IDx-DR" in a form that permits the AI to independently output diagnostic results without the involvement of a physician's judgment. This was the first case in which a regulatory authority officially recognized that "AI may perform medical judgments without a human expert in a specific domain."[7]

1.1.2 Autonomous Driving: AI Safer Than Human Drivers

In the field of autonomous driving, the safety of machine-operated vehicles has recently come to significantly exceed that of human-driven vehicles.

Waymo, a subsidiary of Google, has continuously published peer-reviewed papers on the safety of its autonomous vehicles. According to the latest analysis, based on cumulative public road driving data of over 127 million miles (approximately 200

---

[5] Ardila, D., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25, 954–961.

[6] McKinney, S.M., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577, 89–94.

[7] Abràmoff, M.D., et al. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*, 1, 39.

million kilometers, or roughly 5,000 trips around the Earth) as of 2025, Waymo's autonomous vehicles reduced serious injury crashes by 90%, all injury crashes by 81%, and pedestrian injury crashes by 92% compared with human drivers.[8] In other words, under certain operational conditions, autonomous driving systems can operate vehicles far more safely than human drivers.

1.1.3 The Illusion of "Human Monitoring"

Given these findings—that AI records superior performance to the average qualified professional in high-risk domains subject to regulation—how should rules be designed?

The simplest solution would be to maintain existing regulations while permitting qualified humans to use AI as an assistive tool—that is, a model in which humans monitor AI behavior. Under this approach, the superior judgment capabilities of AI could be fully leveraged, while in the unlikely event that the AI behaves inappropriately, a human could intervene to prevent an accident.

This approach, which appears effective at first glance, in fact harbors a significant problem: when humans and AI collaborate, performance can actually decline.

A meta-analysis of 106 experimental studies on human-AI collaboration (involving over 16,000 subjects in total) found that human-AI collaboration did not yield a "synergy effect" surpassing either humans alone or AI alone. In particular, when AI possessed superior capabilities to humans, human participation substantially degraded performance.[9]

One mechanism underlying this phenomenon is "automation bias." A 2023 study showed that when AI presented incorrect diagnoses, the accuracy of less experienced physicians plummeted to 19.8%, and even experienced physicians' accuracy fell to 45.5%.[10] When AI makes errors, physicians follow suit—and multiple studies have shown that this tendency cannot be prevented through training or awareness-raising.

These findings demonstrate that the intuitive assumption of "safety through human double-checking" does not hold, at least under certain conditions. When humans "second-guess" a capable AI, outcomes worsen.

---

[8] Kusano, K.D., et al. (2025). Comparison of Waymo Rider-Only Crash Rates by Crash Type to Human Benchmarks at 56.7 Million Miles. *Traffic Injury Prevention*, 26(sup1), S8–S20.

[9] Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8, 2293-2303.

[10] Dratsch, T. et al. (2023). Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology*, 307(4).

In the context of autonomous driving, there are also many situations in which humans simply lack the capacity to respond to sudden danger.

Under these conditions, continuing to adhere to the conventional regulatory approach by mandating human monitoring of AI risks not only reducing safety but also imposing disproportionate and ineffective responsibility on the monitoring human, potentially perpetuating a "formalized safety" that fails to serve society's overall interests.

1.2 The Need for Institutional Development Concerning UHAI

1.2.1 What Is UHAI?

Based on the foregoing concerns, this report uses the concept of UHAI—Unsupervised and High-risk Activities AI: AI that autonomously performs high-risk activities without direct human oversight—to examine the appropriate forms of certification and legal frameworks. This section first clarifies the concept of UHAI.

The first defining characteristic of UHAI is that it engages in "high-risk activities." High-risk activities refer to those in which the AI's decisions and actions can directly and significantly affect legally protected interests of citizens, such as life, bodily integrity, property, or fundamental human rights. Specific examples include Level 4 and above autonomous vehicles on public roads, medical devices that autonomously execute diagnoses or treatment plans, algorithmic trading systems that affect financial market stability, and AI systems for the automated control of critical infrastructure. While views on what constitutes "high-risk" vary across countries and cultures (for instance, the EU AI Act designates remote biometric identification and scoring by AI as high-risk domains in addition to existing regulated areas[11]), this report does not delve into that question. As a general principle, decisions subject to safety regulation may be understood as being socially recognized as "high-risk."

The second defining characteristic of UHAI is that AI autonomously performs such high-risk decisions and activities without direct human oversight. Even AI systems that engage in high-risk behavior do not qualify as UHAI if they serve merely as human assistive tools—for example, diagnostic support systems where an AI presents diagnostic candidates but a physician makes the final determination, or Level 2 autonomous driving technology where a human driver continuously monitors the surroundings and can intervene in emergencies.

In essence, UHAI involves the transfer of the "operator" role in the operation of high-risk systems from humans to AI systems. This transition extends beyond mere

[11] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (AI Act), Annex III (High-Risk AI Systems Referred to in Article 6(2)).

improvement of operational efficiency; it has the potential to achieve levels of safety that were previously unattainable through human operation alone. That is, by legally permitting the social implementation of UHAI under certain conditions, it becomes possible to harness AI's latent potential beyond the limits of human cognition and capability, thereby delivering greater safety and benefits to society.

1.2.2 The Barriers of "Evaluation," "Certification," and "Legal Frameworks" Confronting UHAI Implementation

However, there are barriers that must be overcome to achieve the social implementation of UHAI.

First, since UHAI deals with high-risk domains, there must be "evaluation criteria" for determining what requirements an AI must satisfy to be deemed trustworthy. However, the complex characteristics of AI—including its high degree of autonomy and the unpredictability and inexplicability (black-box nature) arising from complex machine learning—make it extremely difficult for consumers, regulators, and even developers themselves to fully assess its safety and reliability. When it is practically impossible to confirm safety by evaluating the algorithm itself, the question of what criteria should form the basis of trust becomes a technical challenge.

Second, for UHAI that operates in high-risk domains, it is often insufficient for the developer alone to inspect whether a system meets these criteria; evaluation and "certification" by a specialized third party may be required. Certification refers to the process by which an independent body evaluates whether a target product, process, service, or system meets specified requirements (conformity assessment) and provides a document (certificate) attesting to this. Given the systemic complexity of UHAI, it is extremely difficult for a third party other than the developer to perform such evaluation and assurance. Accordingly, if a UHAI certification system is to be established, the questions of who should certify, what should be certified, when, and how also become pertinent.

Third, there is also the question of how UHAI should be treated in relation to "legal frameworks." Current regulations presuppose, in principle, that even when AI systems are used, they are employed within the scope of supporting human judgment—that is, the ultimate responsible party remains human. For example, the EU AI Act mandates human oversight for high-risk AI systems,[12] and the operation of UHAI is not, in principle, permitted. In Japan as well, in the medical field, a Ministry of Health, Labour and Welfare administrative notice (I-Sei-I-Hatsu No. 1219-1, December 19, 2018;

[12] Regulation (EU) 2024/1689 (AI Act), Article 14 (Human Oversight).

Iseihatsu dai 1219-gō dai 1)[13] explicitly states the principle that "the actor responsible for diagnosis, treatment, and other acts is the physician, and the physician bears ultimate responsibility for the final judgment," even when AI-based diagnostic support is employed. For autonomous vehicles, although the 2022 amendment to the Road Traffic Act (*Dōro Kōtsū Hō*)[14] makes "Specified Autonomous Driving" (*Tokutei Jidō Unkō*) equivalent to Level 4 legally possible, the authorization requires the deployment of a "Specified Autonomous Driving Supervisor" (*Tokutei Jidō Unkō Shunin-sha*) who performs remote monitoring, and a mode of operation in which AI operates entirely autonomously without any human involvement is not contemplated. As such, under existing legal frameworks, significant institutional barriers remain for the implementation of UHAI, which relies solely on AI's autonomous judgment.

1.3 Technical Characteristics of UHAI as Premises of This Report

This report aims to examine the appropriate standards of trustworthiness for the social implementation of UHAI in light of the above challenges, and to describe the desirable forms of certification and legal frameworks for UHAI.

In doing so, this report focuses on the following three technical characteristics of UHAI. Each is a fundamental technical attribute inherent to UHAI that has implications for all three dimensions: evaluation, certification, and legal frameworks.

1.3.1 Black-box Nature

AI systems that employ deep learning derive their decisions through nonlinear interactions among a vast number of parameters, rendering their decision-making processes an opaque "black box" that is uninterpretable to humans, including the developers themselves. This means that conventional approaches of logically verifying the correctness of internal operations based on specifications—as applied to traditional industrial products—are, in principle, infeasible for AI.

1.3.2 Dynamic Mutability

AI models may change their behavior during operation through continuous learning, fluctuations in data inputs, and changes in the external environment. Unlike planned software updates, this dynamic mutability can occur continuously and nonlinearly, making it impossible to guarantee that a safety assessment conducted at a given point

---

[13] Ministry of Health, Labour and Welfare, Director of Medical Affairs Division, Medical Affairs Bureau, "Concerning the Relationship Between the Use of Programs That Provide Diagnostic, Therapeutic, and Other Support Using Artificial Intelligence (AI) and the Provisions of Article 17 of the Medical Practitioners' Act" (I-Sei-I-Hatsu No. 1219-1, December 19, 2018).

[14] Act Partially Amending the Road Traffic Act (Act No. 32 of 2022). Provisions relating to Specified Autonomous Driving (*Tokutei Jidō Unkō*) are found at Article 75-12 and subsequent articles.

in time (a snapshot assessment) will remain valid in the future. Furthermore, since UHAI operates without a human supervisor, there is a risk that performance degradation and unexpected behavioral changes accompanying dynamic mutability may go undetected for extended periods.

1.3.3 Difficulty in Goal-Setting

Since UHAI is a system fundamentally based on probability and statistics, it is impossible in principle to guarantee "absolute safety," and a certain degree of inherent risk is unavoidable. It is therefore necessary for society to define "what level of safety should be achieved," yet there is no absolute correct answer regarding this level; acceptability varies depending on cultural context and societal needs. Moreover, the very methodology for measuring and evaluating qualitative requirements such as "fairness (prevention of discriminatory behavior)" and "explainability (presentation of reasoning)," in addition to physical safety, remains undeveloped, making it inherently difficult to establish the "yardstick" for evaluation itself.

1.4 Structure of This Report

Based on the technical characteristics of UHAI described above, this report proceeds as follows. Chapter 2 analyzes the gap between the nature of UHAI and existing institutional frameworks from three perspectives: evaluation, certification, and legal frameworks. Chapter 3, seeking clues to bridge these gaps, examines as case studies how regulatory models in major countries address the challenges posed by UHAI, focusing on two domains: AI medical devices (SaMD) and autonomous driving (AV). Chapter 4, drawing on these analyses, identifies key issues for the construction of evaluation criteria, the introduction of new certification systems, and the redesign of legal frameworks.

2. Three Institutional Barriers Impeding the Implementation of UHAI

The following analysis examines the gap between the nature of UHAI and existing institutional frameworks from three perspectives: evaluation (2.1), certification (2.2), and legal frameworks (2.3).

2.1 Challenges in the Evaluation of UHAI

The fundamental factor that makes the certification of UHAI so difficult lies in the inherent difficulty of establishing evaluation criteria—that is, the "yardstick" for determining what constitutes a "pass."

2.1.1 The Difficulty of Evaluating AI Products

Conventional certification of industrial products (e.g., electrical equipment) has generally relied on a "specification conformity" approach, which tests whether a product conforms to predefined specifications and performance standards. However, applying this approach to UHAI—which, as discussed in Section 1.3, exhibits the characteristics of "black-box nature," "dynamic mutability," and "difficulty in goal-setting"—encounters the following limitations.

First, with respect to the black-box nature, the decision-making processes of UHAI are uninterpretable to humans, making it difficult to logically verify whether the system operates in accordance with its specifications. Consequently, exhaustive testing is impossible, and the safety of the product cannot be fully guaranteed.

Second, with respect to dynamic mutability, the results of an evaluation conducted at a given point in time—a "snapshot"—do not guarantee the safety of the UHAI throughout its operational life.

Third, with respect to the difficulty in goal-setting, the very formulation of evaluation indicators and evaluation methodologies for assessing the reliability of UHAI is inherently challenging.

2.1.2 Limitations of Evaluating Operators (Organizations)

As a complement to the limitations of product evaluation, there is an approach that evaluates not the product itself but rather the organization that develops and operates it. Management system certification (organizational certification) based on ISO/IEC 42001 falls into this category. This approach evaluates the "organizational governance structures" that enable organizations to identify and analyze AI risks and pursue continuous improvement through PDCA cycles and similar mechanisms, and it plays an extremely important role as a foundation for ensuring AI trustworthiness.

However, in the context of UHAI certification, this organizational certification approach, when used "in isolation," has the following limitations. Management system certification only guarantees that the organization is "developing and operating in accordance with appropriate processes"; it does not directly guarantee that "individual UHAI products produced through those processes will behave safely under specific conditions." In particular, for UHAI that makes autonomous decisions without human intervention, even if the organization is in excellent health, it is impossible to completely eliminate the risk that the black-box product itself may exhibit unexpected behavior (such as hallucinations or accidents).

Therefore, to underpin the trustworthiness of UHAI, it is essential not only to evaluate the management capabilities of the organization but also to ascertain that the product-specific risk measures under its management are functioning effectively.

Thus, neither a product certification approach nor a management system certification approach alone is sufficient to address the risk characteristics inherent in UHAI—namely, its black-box nature, dynamic mutability, and difficulty in goal and evaluation methodology setting.

2.2 Challenges in Granting Certification for UHAI

Even if evaluation criteria for UHAI could be formulated, the construction of a certification system to determine who should certify, when, and how still faces formidable hurdles.

First, there is the problem of information asymmetry and the limitations of third-party evaluation. The party best positioned to understand the risks and behavior of an AI system is the developer (provider) itself, and it is extremely difficult for external certification bodies to accurately assess the internals of a black-box AI within the constraints of limited information and time.

Second, there is the problem of certification timing (time-to-market). The pace of AI technological evolution and model update frequency is extremely rapid. Under a conventional system design that requires months for certification, by the time certification is obtained, the model may have already been updated or become technologically obsolete.

Third, there is the difficulty of scoping the certification target. UHAI does not exist in isolation; it operates in complex interaction with hardware (e.g., vehicles, medical devices), cloud environments, and external services connected via APIs. Determining the boundaries of the certification scope within this "system of systems" is a technically and legally unresolved challenge.

2.3 Challenges in the Legal Frameworks for UHAI

Even if some form of evaluation criteria were established and a certification system put in place, unresolved problems would remain. To what extent can pre-market regulation control UHAI? If a regulatory framework is established, can it be effectively enforced? And if UHAI actually causes harm, who bears legal responsibility and how? These are structural challenges inherent to legal frameworks that exist independently of the evaluation and certification issues.

The Japanese government, in response to the rapid development of AI technology, has pursued a flexible governance model based on the concept of agile governance rather than rigid ex ante regulation.[15] More recently, the interpretation and application of

[15] Ministry of Economy, Trade and Industry, Study Group on New Governance Models in Society 5.0, "GOVERNANCE INNOVATION—Re-Design of Law and Architecture for Realizing Society 5.0" (2020);

civil liability in the context of AI utilization have also been examined within the existing legal framework.[16] These policy efforts provide an important foundation for the broader development of legal frameworks for AI.

The following sections examine this issue from three dimensions—pre-market regulation (2.3.1), enforcement (2.3.2), and liability regime (2.3.3)—while taking into account the directions that the Japanese government has articulated.

2.3.1 Challenges Concerning Pre-market Regulation

Pre-market regulation, in any jurisdiction, is premised on the assumption that the specifications of the subject can be determined in advance and that regulatory authorities can examine their content. However, the three characteristics of UHAI—black-box nature, dynamic mutability, and difficulty in goal and evaluation methodology setting—expose the limits of this framework.

First, the black-box nature undermines the premises of the authorization system. Authorization presupposes that the applicant can articulate the specifications of the subject system and explain its safety. However, the decision-making process of UHAI is difficult to explain even for the developers themselves, and the very act of fixing specifications may not be possible. For UHAI, the problem is not merely that specifications are unclear. Rather, because the action space itself can be generatively expanded during operation, the very concept of specifiable specifications in advance may not hold. For AI in which humans are involved in operation, the safety valve of "ultimately, a human will confirm and judge" can function; for UHAI, this pathway does not exist. There is no legal basis under current systems for granting authorization to a subject whose specifications cannot be fixed.

Second, whereas conventional authorization has been a conformity judgment regarding the state of a system at a specific point in time, the dynamic mutability of UHAI severely limits its validity. This is not simply a technical problem solvable by "increasing update frequency." Because changes in UHAI occur continuously and nonlinearly, the structural premise underlying authorization—namely, "fixing the state at a given point in time"—becomes difficult to sustain as an institutional assumption. For AI with human oversight, the operator may notice that "recent outputs seem off"; for UHAI, since there is no such monitoring agent, there is a risk

---

"GOVERNANCE INNOVATION Ver.2: Toward the Design and Implementation of Agile Governance" (2021); "GOVERNANCE INNOVATION Ver.3: Overview and Current Status of Agile Governance" (2022).

[16] Ministry of Economy, Trade and Industry, "Study Group on Civil Liability in AI Utilization" (August 2025–). The study group aims to formulate guidelines on the interpretation and application of civil liability in AI utilization, organizing the interpretation and application of existing law (tort, product liability, breach of contract, etc.).

that deviations will go undetected for extended periods. Current authorization systems have no answer to the question of "how long does an authorization remain valid?"

Third, the difficulty in goal-setting impedes the formulation of regulatory standards. As discussed in Section 2.1, the very safety threshold of UHAI is difficult to define. The effectiveness of guidelines presupposes clear standards to be observed, but the capabilities of UHAI are cross-domain and context-dependent, and do not lend themselves to uniform standards. Moreover, the inability to supplement gaps in standards through ad hoc human judgment compounds the problem.

As described above, the convergence of three problems—specification indeterminacy, loss of temporal validity, and absence of standards—causes the premises of pre-market regulation to structurally break down. Furthermore, because UHAI lacks a "last-resort human safety valve," these problems become all the more acute.

2.3.2 Challenges Concerning Enforcement

The enforcement of laws against UHAI also presents formidable challenges.

First, the black-box nature undermines the effectiveness of oversight. For AI under human supervision, enforcement authorities can indirectly monitor system status through the channel of operators detecting anomalies and reporting them. This quasi-whistleblowing pathway does not exist for UHAI, making external monitoring the sole means—yet external monitoring itself is constrained by the black-box nature. Conventional enforcement measures such as on-site inspections and demands for reports presuppose that the internal state of the subject is at least partially visible. The internal state of UHAI is opaque even to the developers, and the means available to enforcement authorities to ascertain "what is happening" from the outside are extremely limited. Monitoring only observable inputs and outputs cannot assess the appropriateness of internal decision-making processes.

Second, dynamic mutability makes violation determination difficult. The determination of a "violation" requires identification of the precise moment of violation, but because UHAI changes gradually and continuously, there may be no clear answer to the question of "when did the system become non-compliant." Furthermore, because there is no human supervisor, there is a structural risk that non-compliant conditions may accumulate over extended periods without detection. This can lead to situations in which, by the time a problem becomes apparent, harm has already spread.

Third, the difficulty in goal-setting leaves enforcement design without a foundation. What should be monitored, by what indicators should judgments be made, and at what frequency should inspections be conducted—all parameters necessary for the concrete design of enforcement remain undefined. Moreover, given that enforcement

agencies themselves do not currently possess sufficient capacity for the technical evaluation of UHAI, effective monitoring is extremely difficult in practice.

Thus, even if regulations are established through pre-market mechanisms, the means to effectively enforce them have not kept pace. The gap between regulation and enforcement fundamentally undermines the effectiveness of the legal system.

2.3.3 Challenges Concerning Liability

The question of how to address ex post liability for harm caused by UHAI is also of critical importance within the legal framework for UHAI. And, as with the pre-market regulation and enforcement issues discussed above, the three structural characteristics of UHAI pose various difficulties for the allocation of liability.

First, the black-box nature makes causal proof difficult. The establishment of tort liability requires the victim to prove the causal relationship between the injurious act and the harm, but in the case of UHAI, it is difficult even for experts to examine the AI's decision-making process. Furthermore, for AI under human supervision, the issue can ultimately be framed as one of causation between human error and the outcome, but no such legal fiction is available for UHAI. The chain of causation is entirely sealed within the black box. In the EU, a proposal to address this problem by creating a presumption of causation for high-risk AI systems (the AI Liability Directive) was put forward, but the draft directive failed to gain consensus among Member States and was formally withdrawn in October 2025.[17]

Second, dynamic mutability creates inconsistencies with existing concepts of "defect." The Product Liability Act (*Seizōbutsu Sekinin Hō*) requires a "defect at the time of delivery" as a prerequisite.[18] However, because UHAI continues to learn and change after shipment, even if there was no problem at the time of delivery, performance may degrade or unexpected decision-making tendencies may be acquired during

---

[17] European Commission, Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final (September 2022). The draft directive, which included provisions for presumption of causation and evidence disclosure orders for high-risk AI, failed to gain consensus among Member States and was announced for withdrawal in the European Commission's 2025 Work Programme (COM(2025) 45), Annex IV, and formally withdrawn via Official Journal C/2025/5423 (October 6, 2025).

[18] Article 2, Paragraph 2 of the Product Liability Act (*Seizōbutsu Sekinin Hō*) defines "defect" as "a lack of safety that the product ordinarily should provide, taking into account the characteristics of the product, the manner in which it is ordinarily foreseeable to be used, the time at which the manufacturer, etc. delivered the product, and other circumstances relating to the product." The fact that "the time of delivery" serves as the reference point is the issue in relation to UHAI.

operation. Current law provides no clear answer to the question of "at what point in time should the state be used as the baseline for judging a defect."

Third, the difficulty in goal-setting makes the delineation of the scope of liability difficult. The determination of negligence requires establishing "foreseeability" and "the duty to avoid the outcome," but as discussed in Section 2.2, the very methodology for evaluating UHAI safety is undeveloped, making it difficult to determine what should have been foreseen and what outcomes should have been avoided. Furthermore, since an AI's supply chain involves various actors including developers, providers, and users, it is extremely difficult to define the scope of foreseeability and the duty to avoid outcomes for each of these actors.

2.3.4 Summary

What has become clear through this analysis is that UHAI creates structural challenges across three dimensions of existing legal frameworks. What UHAI calls into question is not a shortage of regulation, but rather the fundamental problem of the collapse of the premise that risk is controlled by humans. In pre-market regulation, the premises of regulation itself do not hold due to specification indeterminacy, the absence of standards, and the loss of temporal validity. In enforcement, even if regulations exist, the means to make them effective are lacking due to the invisibility of system internals, the difficulty of identifying when violations occur, and the challenge of formulating enforcement standards. In the liability regime, the pathway for victim relief becomes unstable due to the difficulty of proving causation, the inconsistency between the defect concept and the dynamic nature of AI, and the difficulty of delineating the scope of liability.

The challenges across these three dimensions do not exist in isolation but are mutually reinforcing. If pre-market regulation fails to function, the boundary between lawful and unlawful conduct remains unclear, making it impossible to define the scope of enforcement. If enforcement fails to function, violations go uncorrected and harm occurs, with the entire burden transferred to the liability regime. But the liability regime also malfunctions due to the difficulty of proving causation and delineating the scope of liability, thereby closing the pathway for victim relief.

This vicious cycle within the legal system is also interlinked with the evaluation and certification challenges examined in Sections 2.1 and 2.2. If evaluation criteria are indeterminate, certification becomes difficult; if certification is difficult, legal safe harbors cannot be established, and liability risks expand further.

To break this vicious cycle, it is important to understand what efforts are currently being made, both domestically and internationally, concerning legal frameworks for advanced AI. Accordingly, Chapter 3 presents case studies on medical AI and autonomous driving.

3. Current Institutional Achievements and Challenges (Case Studies)

As noted in the preceding chapter, UHAI brings to the fore three fundamental challenges: "black-box nature," "dynamic mutability," and "difficulty in goal and evaluation methodology setting."

This chapter takes up two representative high-risk domains—medical devices (SaMD) and autonomous driving (AV)—and closely examines how the regulatory models of major countries confront these three challenges.

As a method of analysis, Sections 3.1 and 3.2 first examine the institutional responses in each case along the three technical characteristics: black-box nature, dynamic mutability, and difficulty in goal and evaluation methodology setting. This approach enables comparison of how different countries have adopted different approaches to the same technical challenges. Section 3.3 then "translates" these findings into the framework of the three institutional challenges raised in Chapter 2—evaluation, certification, and legal frameworks—thereby bridging to the recommendations in Chapter 4.

3.1 Case Study 1: Certification of AI-enabled Medical Devices

As noted at the outset of this report, the domain of Software as a Medical Device (SaMD)—that is, software that functions as a medical device by itself making diagnostic or therapeutic decisions, rather than being embedded in hardware—is the field where AI utilization first became a subject of regulatory concern. As of 2025, the FDA has authorized more than 1,300 cumulative AI-enabled medical devices, with 331 approved in a single year.[19]

This section analyzes how SaMD regulation in Japan and the United States addresses the three challenges of UHAI, and also examines the relationship between organizational certification and product certification.

3.1.1 Overview of the Regulatory Framework

Japan: The IDATEN Framework

The approval review of AI medical devices in Japan is based on the same framework as conventional medical device regulation. Rather than directly verifying internal structures or mechanisms, safety and efficacy are assessed through a combination of

[19] FDA, Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices . https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

clinical performance testing (evaluating diagnostic accuracy using clinical data) and non-clinical performance testing (performance evaluation using bench tests and known datasets)—fundamentally, a behavioral verification approach examining "what outputs are returned for given inputs." Under the implicit assumption that "internal structures are stable," this framework functioned adequately for conventional medical devices. However, for AI, whose internals are not only a black box but whose behavior can also change through learning, the same evaluation framework cannot guarantee safety after such changes—a structural problem that prompted the introduction of the IDATEN framework through the 2019 amendment to the Pharmaceutical and Medical Device Act (*Yakki Hō*).

The Post-Approval Change Management Protocol (PACMP) system introduced by the 2019 amendment to the Pharmaceutical and Medical Device Act—commonly known as IDATEN (Improvement Design within Approval for Timely Evaluation and Notice)—is a framework for expediting change management of SaMD, including AI. Developers submit a plan for future algorithm changes at the time of initial approval, and after confirmation by the Pharmaceuticals and Medical Devices Agency (PMDA), changes within the planned scope can be implemented with only a minor change notification. The essence of this system lies in treating AI algorithm changes as "changes within a pre-planned scope," thereby eliminating the need to restart approval procedures from scratch for each change. However, the framework's actual utilization remains limited; for therapeutic SaMDs in particular, clinical trials may be required not only for changes to the core mechanism but even for the addition of user support functions, and questions have been raised about its practical effectiveness given the balance between plan preparation and review costs on the one hand and procedural streamlining benefits on the other.[20]

United States: The FDA PCCP Framework

In December 2024, the US FDA issued final guidance on Predetermined Change Control Plans (PCCPs) for AI-enabled medical devices.[21] Under the PCCP framework, manufacturers present at the time of initial submission: (i) a Description of Modifications; (ii) a Modification Protocol that specifies the methodology for verifying

---

[20] Ministry of Health, Labour and Welfare, "Guidance for Appropriate and Rapid Approval and Development Taking into Account the Characteristics of Software as a Medical Device (SaMD)," 2nd Edition (June 5, 2024) (https://www.pmda.go.jp/files/000269089.pdf). For discussion of practical difficulties in utilizing the IDATEN framework, see "Unnecessary Regulation of SaMDs Risks Dampening Development Motivation—Current Status and Challenges of SaMDs, Vol. 3," m3.com (March 14, 2025) (https://www.m3.com/news/open/iryoishin/1261551).

[21] FDA, Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions: Final Guidance (December 2024).

changes; and (iii) an Impact Assessment. The FDA then reviews and approves the overall adequacy of the process. Changes within the approved scope can be implemented without additional submissions. Furthermore, the PCCP operates in conjunction with the Total Product Life Cycle (TPLC) approach. That is, continuous monitoring of Real-World Performance (RWP)—performance evaluation based on actual clinical use data—is required, and a mechanism for verifying in the marketplace whether post-change products function as expected constitutes an indispensable component of the system.

To understand the institutional significance of the PCCP, it is important to note that it is not merely a "simplification of change procedures." Under the PCCP, the subject of FDA review shifts from "the post-change product" to "the process that produces the change itself." Under conventional change review, the FDA individually examined the post-change product for each modification. Under the PCCP, the FDA reviews the design of the process—"how changes will be verified and by what criteria pass/fail will be determined"—and if the process is judged adequate, individual change outcomes that pass through the process do not require additional review. Specifically, the PCCP review assesses whether the data requirements, verification methods, and acceptance criteria contained in the manufacturer's change protocol are appropriate in light of the risk profile of the specific product (evaluation of product-specific risk management measures). At the same time, the organization's capability to execute and continually comply with that protocol—that is, its quality management system, change management processes, and post-market surveillance infrastructure—also falls within the scope of FDA review (evaluation of organizational change management capability). This dual evaluation structure was not established by the PCCP alone but has been shaped in conjunction with the development of Good Machine Learning Practice (GMLP) (standardization of organizational-level good practices) and the TPLC approach (management throughout the product lifecycle).

The PCCP is a framework informed by the lessons of the "Pre-Cert (Software Precertification) Program" that the FDA piloted in the late 2010s. Pre-Cert was an ambitious concept that sought to certify the quality management capabilities of development organizations in lieu of individual product reviews, with products from certified organizations exempted from pre-market review—in other words, an attempt to demonstrate whether "excellent organizations produce safe products." However, the program was ultimately discontinued. The reasons were multifaceted. First, it was found that even for relatively low-risk Class 2 medical devices, evaluating organizational quality culture through KPIs alone could not cover the risks specific to individual products. Second, the FDA lacked sufficient legal authority to collect from companies the information necessary for certification. Third, the pilot program's

participating companies were disproportionately large corporations such as Apple and Johnson & Johnson, drawing criticism from startups regarding unfairness.[22]

The lesson from the discontinuation of Pre-Cert is that organizational certification alone cannot guarantee product safety. However, it does not follow that organizational certification methods were entirely without merit. The partial achievements of Pre-Cert led to the development in 2021 of the "Good Machine Learning Practice (GMLP) for Medical Device Development" by the FDA, the UK MHRA, and Health Canada,[23] and subsequently to the finalization of the PCCP in 2024. The current PCCP framework thus represents the convergence toward combining pre-assessment of organizational change management capabilities with review of product-specific risk management measures, embodying the direction of "staged integration of product certification and organizational quality management."

United States: The IDx-DR Precedent for Autonomous AI Approval

In 2018, the FDA approved the diabetic retinopathy diagnostic AI "IDx-DR" (now LumineticsCore) through the De Novo pathway—that is, the pathway for novel medical devices with no existing predicate, which creates a new regulatory classification for authorization—in a form that permits the AI to independently output diagnostic results without physician involvement.[24] This was the first case in which a regulatory authority officially recognized that "AI may perform medical judgments without a human expert in a specific domain," and it holds extremely significant implications as a precedent for the social implementation of UHAI.

The approval structure of IDx-DR contains several features instructive for the institutional design of UHAI. First, the design places responsibility for the final output on the AI developer, clearly delineating the responsible entity within the product certification framework. Second, training requirements are imposed not only on the developer but also on the facilities that use the device, creating a structure that integrates product-side risk management measures with requirements for the organizational preparedness of the user institution. Third, the approval is limited to specific conditions of use (screening for diabetic retinopathy in primary care settings), and the "limitation of the scope of application"—analogous to the delineation of an Operational Design Domain (ODD)—serves as a prerequisite for ensuring safety.

---

[22] FDA, The Software Precertification (Pre-Cert) Pilot Program: Tailored Total Product Lifecycle Approaches and Key Findings (September 2022).

[23] FDA, Health Canada & MHRA, Good Machine Learning Practice for Medical Device Development: Guiding Principles (October 2021).

[24] Abràmoff, M.D., et al. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine, 1, 39. Same source as note 7.

3.1.2 Response to the Three Technical Challenges

Based on the institutional frameworks described above, the following sections organize the implications that Japanese and US medical device AI regulation offers for addressing the three technical challenges of UHAI: black-box nature, dynamic mutability, and difficulty in goal-setting.

3.1.2.1 Response to the Black-box Problem

In response to the opacity of AI medical device decision-making, Japanese and US regulatory frameworks adopt different approaches, yet both share a common feature: they abandon direct elucidation of AI internals and instead seek to ensure safety through indirect methods.

(i) Product-level response: "behavioral evaluation" through input-output verification

As described in Section 3.1.1, Japan's IDATEN framework fixes (locks) the post-change algorithm and then verifies input-output relationships through clinical and non-clinical performance testing. Rather than elucidating the "internals" of the black box, this is an approach that verifies "behavior." However, particularly for therapeutic SaMDs, the high cost of evaluation acts as a factor impeding rapid system improvement.[25]

The US PCCP framework takes this behavioral verification one step further, making the subject of review not the outcome of individual changes but the adequacy of the process that produces changes. If the "change protocol"—including data requirements for retraining, verification methods, and acceptance criteria—has been approved in advance, the results that pass through that process are institutionally tolerated even though they are a black box. This represents an important conceptual shift—not "resolving" the black-box nature but "incorporating" it within the institutional framework.

(ii) Organizational-level response: supplementing the limitations of product evaluation with "evaluation of organizational management capabilities"

However, product input-output verification alone cannot sufficiently ensure the safety of black-box AI. Input-output verification is merely a test result under specific conditions, and it is impossible in principle to cover every situation that may be encountered during actual operation. Accordingly, movements to evaluate "the

[25] Ministry of Health, Labour and Welfare, "Guidance for Appropriate and Rapid Approval and Development Taking into Account the Characteristics of Software as a Medical Device (SaMD)," 2nd Edition (June 5, 2024). See in particular Section III-2 regarding clinical trial requirements for therapeutic SaMDs.

management capabilities of the organization that develops and operates the product" in addition to product evaluation have attracted attention.

As detailed in Section 3.1.1, the FDA's Pre-Cert program attempted to guarantee product safety through organizational evaluation alone and failed, but its lessons have been incorporated into the PCCP framework. Under the PCCP, review of the product's change protocol (product level) and review of the organization's quality management system that executes that protocol (organizational level) are integrated. In the IDx-DR approval structure as well, training requirements for user facilities are imposed alongside limitations on conditions of use, embedding evaluation targeting both the product and the organization within the regulatory framework. In this way, an approach of managing the behavioral risk of black-box products through a two-stage mechanism—product input-output verification and organizational governance—is emerging across both Japan and the United States.

3.1.2.2 Response to Dynamic Mutability

How to ensure the safety of AI that continuously learns and changes is a core challenge of medical device regulation.

Japan's IDATEN framework introduces a flexible mechanism in the form of change plans, yet in practice requires the algorithm to be locked at each change point and a 30-day confirmation period following notification to the PMDA before implementation. This effectively forces the continuous evolution of AI into "incremental frame-by-frame" steps, risking the substantial limitation of the technological advantages of continuous learning.

The US PCCP framework, based on the concept of the TPLC (Total Product Life Cycle) approach, places post-market Real-World Performance (RWP) monitoring at the core of safety assurance. If a protocol has been approved in advance, individual changes can be implemented without additional approval, thereby opening the door to incorporating dynamic mutability within the institutional framework. Nevertheless, actual utilization of the PCCP remains limited. According to FDA medical device databases, a cumulative total of 67 devices have received PCCP authorization, of which 59 (88%) were concentrated in the most recent two years. Of the more than 3,000 510(k) clearances issued in 2024, only 41 (approximately 1%) included a PCCP. These figures indicate that while the institutional direction is becoming established, practical adoption is still in its early stages.[26]

[26] Gardner Law, Streamlining Device Changes with Predetermined Change Control Plans (PCCPs) (2025), citing FDA Medical Device Database.

3.1.2.3 Response to the Difficulty in Goal and Evaluation Methodology Setting

The question of what constitutes "safe" is a problem that lies at the foundation of medical device AI regulation.

Japan's IDATEN framework applies the existing medical device regulatory framework (efficacy and safety) as is, and no evaluation methodology specific to AI safety has been established. Faced with AI that entails probabilistic risk, regulators tend to demand time-consuming clinical trial evidence, with the result that the system risks becoming rigid.

The US PCCP framework addresses the standard-setting problem by requiring manufacturers to present "performance criteria and acceptance criteria" in advance, but the validation of those criteria themselves depends heavily on the developer's self-verification. The FDA supplements this with post-market RWP monitoring, but in domains where the evaluation methodology has not been established, the developer ultimately determines that methodology themselves. This risks ultimately falling back on trust in the organization.

The consequences of this "difficulty of designing output evaluation" are strikingly illustrated by the experience of Germany's Digitale Gesundheitsanwendungen (DiGA) framework. DiGA is a "fast track" system under which physician-prescribable app-type medical devices (such as diabetes management apps and cognitive behavioral therapy apps) receive provisional insurance coverage, with manufacturers required to submit evidence of clinical efficacy within a specified period. The design lowers the entry barrier to promote innovation while verifying efficacy at the exit (continued registration) through post-market evaluation. However, of the 68 products that received fast-track designation, only 12 (less than 20%) demonstrated efficacy and achieved continued registration within the prescribed period.[27] The DiGA experience shows that streamlining and expediting the entry process alone is insufficient; without a system for continuously verifying AI performance based on reliable data after market entry—that is, the public infrastructure for post-market monitoring—the credibility of the entire regulatory framework is undermined.

Furthermore, integrating international standards presents additional difficulties. Revision of ISO 13485, the quality management system standard for medical devices, is not planned until around 2030, and integration with AI-specific standards (such as

[27] Dahlhausen, F., et al. (2022). There's an App for That, but Nobody's Using It: Insights on Digital Health Application Evidence Requirements and Reimbursement in Germany. Digital Health, 8. See also the BfArM DiGA database.

ISO/IEC 42001) is not expected in the near term.[28] In other words, the international framework for institutionally connecting domain-specific certification requirements for medical devices with certification requirements specific to AI systems remains under development.

3.1.3 Summary

The analysis in this section reveals the following directions and limitations in Japanese and US medical device AI regulation. With respect to the black-box nature, both countries have converged on an approach that abandons efforts to elucidate internal structures in favor of input-output verification (Japan's post-lock clinical evaluation and the US protocol review). Furthermore, recognizing that behavioral verification of products alone has its limitations, a "two-stage evaluation" approach that also assesses organizational management capabilities is observed in the PCCP and IDx-DR approval structures. However, as the failure of Pre-Cert demonstrates, the lesson that organizational evaluation alone cannot substitute for product safety is also shared.

With respect to dynamic mutability, Japan's IDATEN framework adopts "incremental frame-by-frame" steps through locking, while the US PCCP framework opens the door to incorporating change within the institutional framework through the TPLC approach and RWP monitoring. However, PCCP utilization accounts for only approximately 1% of all approvals, and practical adoption is still in its early stages.

With respect to the difficulty in goal and evaluation methodology setting, the fundamental challenge common to both countries is that the criteria for "what constitutes safe" remain unestablished. Japan takes a conservative approach relying on clinical trials, while the United States takes a flexible approach that relies on developer self-verification, but neither has succeeded in establishing a safety evaluation methodology specific to AI.

3.2 Case Study 2: Certification of Autonomous Driving Systems (AV)

The autonomous driving domain is the field where the "elimination of human involvement" manifests most prominently in the context of UHAI. When vehicle control transitions from "driving assistance (Level 2)" to "fully autonomous operation (Level 4 and above)," the system ceases to be a mere tool and becomes a "driving

[28] ISO 13485:2016 was judged to require no revision (confirm) in its 2025 systematic review and is expected to remain valid until around 2030. See TÜV NORD, "ISO 13485:2016 reconfirmed—Valid until April 2030" (2025).

agent" that replaces the human.[29] This section analyzes the regulatory models of Japan, the United Kingdom, and the United States (California), examining how each system confronts the three challenges of black-box nature, dynamic mutability, and difficulty in goal and evaluation methodology setting.

3.2.1 Overview of the Regulatory Framework

Japan: The Specified Autonomous Driving Authorization System

The 2022 amendment to the Road Traffic Act (*Dōro Kōtsū Hō*) (effective April 2023) created an authorization system for "Specified Autonomous Driving" (*Tokutei Jidō Unkō*) equivalent to Level 4. Business operators (Specified Autonomous Driving Operators) must submit a Specified Autonomous Driving Plan to the prefectural Public Safety Commission that includes the Operational Design Domain (ODD—the geographic scope, weather conditions, speed, and other parameters within which the autonomous driving system can function safely), the remote monitoring system, and cybersecurity measures, and obtain authorization. The deployment of a Specified Autonomous Driving Supervisor (*Tokutei Jidō Unkō Shunin-sha*) to perform remote monitoring and the installation of a Data Storage System for Automated Driving (DSSAD) are mandatory.

United Kingdom: The Automated Vehicles Act 2024

The Automated Vehicles Act (AVA), enacted in May 2024, created a new category of responsible entity called the Authorized Self-Driving Entity (ASDE, entities bearing legal responsibility as the 'driver' in automated driving systems.).[30] During automated driving mode, persons in the vehicle are released from legal responsibility for driving operations, and the ASDE assumes all obligations. To be authorized as an ASDE, it must be demonstrated that the system can maintain a level of safety "equivalent to or better than a careful and competent human driver" (the C&C standard).

A particularly noteworthy aspect of the UK's ASDE model is its adoption of a two-tier structure comprising vehicle type approval and organizational authorization (ASDE authorization). That is, it constitutes a framework that integrates the safety evaluation of a vehicle as a product with the capability evaluation of the organization that continuously operates and manages that product.[31] To be authorized as an ASDE, the following are required: technical capability in safety (the ability to design, test, and

---

[29] SAE International, J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (Revised April 2021).

[30] Automated Vehicles Act 2024, c. 10 (UK). The Act received Royal Assent on May 20, 2024.

[31] Law Commission of England and Wales & Scottish Law Commission, Automated Vehicles: Joint Report, Law Com No 404, Scot Law Com No 258 (January 2022).

continuously monitor the autonomous driving system); an appropriate corporate governance structure (established decision-making processes for safety matters); financial soundness (sufficient resources to address damage compensation); and an information disclosure framework (organizational mechanisms for honest reporting to the authorities). Furthermore, after obtaining authorization, the ASDE bears an ongoing obligation to continuously monitor the safety of the system and promptly report any detected problems to the authorities.

From the perspective of the liability regime, a further noteworthy feature of the AVA is the structure of civil liability for accidents during automated driving mode. The AVA adopted a structure whereby, in the event of an accident caused by an autonomous vehicle, persons inside the vehicle are released from legal responsibility for driving operations, while the ASDE bears responsibility if it has failed to meet the safety standard of "equivalent to or better than a careful and competent driver." This fundamentally transforms the legal structure under which conventional automobile accident liability was attributed to the individual "driver," unifying the safety assurance obligation in the organization (ASDE). This transfer of liability is one legislative answer to the question of "who should be the ultimate responsible party" in the context of UHAI.

At the same time, the AVA also incorporates a mechanism whereby having constructed and maintained an appropriate safety management system as an organization can serve as a basis for mitigating the ASDE's liability. If the ASDE has fulfilled authorization requirements, complied with the Duty of Candor, and conducted continuous safety monitoring, yet an accident still occurs, the satisfaction of these organizational requirements may function as a form of safe harbor. In the criminal law context, this is structured as a "due diligence defence"—if the ASDE can demonstrate that it took the measures reasonably expected for ensuring safety, it may be released from criminal liability. This mechanism not only transfers liability to the organization but also incentivizes the construction of appropriate governance structures themselves. It may be characterized as an institutionalization of the "compliance as defence" structure.[32]

California (US): The Deployment Permit System

The California Department of Motor Vehicles (DMV) operates a staged permit system (test permit → driverless test permit → deployment permit).[33] At the federal level, under the self-certification system of the National Highway Traffic Safety Administration (NHTSA), manufacturers submit a Voluntary Safety Self-Assessment

[32] Automated Vehicles Act 2024, Part 1, section 10; Law Commission Joint Report (2022) and following.

[33] Cal. Veh. Code §§ 38750–38755; Cal. Code Regs. tit. 13, §§ 227–228.

(VSSA). The development of specific technical standards remains limited, with data disclosure obligations and post-market monitoring serving as the primary mechanisms.

3.2.2 Response to the Three Technical Challenges

3.2.2.1 Response to the Black-box Problem

In response to the opacity of autonomous driving AI decision-making, the three countries adopt different approaches.

Japan mandates the recording of driving data through the DSSAD and, in the development of vehicle safety standards, is developing a methodology that evaluates through scenario-based testing whether safety levels equivalent to or better than a "competent and careful human driver" have been achieved (the SAKURA Project).[34] Rather than directly verifying AI's internal decision-making processes, this is an approach that determines safety based on outcomes (logs and scenario test results). However, this approach relies solely on outcomes and lacks the means to directly evaluate AI's internal logic. Consequently, the institutional design deploys a Specified Autonomous Driving Supervisor to perform remote monitoring, supplementing AI's uncertainty through human involvement—that is, it does not reach the level of UHAI.

The United Kingdom, through the ASDE framework, circumvents the black-box problem via "legal fiction." Instead of asking "why" the AI reached a particular decision, driving outcomes themselves are attributed to the ASDE as its "acts." This means that even if AI internals are opaque, the ASDE bears statutory liability when C&C standards are breached. By translating the technical black box into organizational responsibility, it provides a clear legal answer. Accordingly, stringent requirements are imposed to ensure that the ASDE can continuously bear the responsibility for safety assurance.

California secures transparency through the VSSA self-assessment process and accident data disclosure obligations. The emphasis is placed on the publication of driving performance data rather than the elucidation of internal decision-making processes.

3.2.2.2 Response to Dynamic Mutability

Managing systems that continuously change through OTA (Over-the-Air) updates is a core challenge of autonomous driving regulation.

---

[34] SAKURA Project. https://www.sakura-prj.go.jp/jp/

Japan's system is premised on snapshot assessment based on a specific software version and ODD combination. Because change notifications or re-examinations may be required for each update, a structural challenge arises between the dynamic evolution of AI and the static character of administrative authorization.

The United Kingdom adopts a dynamic management model centered on the ASDE's "Duty of Candor"—the obligation to promptly report to the authorities, without concealment, any safety risks or defects that come to its attention—and continuous safety monitoring. Particularly noteworthy is the requirement to designate a named senior manager as the person responsible for information provision, with criminal liability imposed on the individual for the provision of false information or intentional concealment. This institutional design intentionally creates a structural tension between the corporation and the individual. Even if a corporation were to attempt to systematically conceal safety problems, the designated senior manager faces the risk of personal criminal liability for participating in the concealment, thereby creating an incentive for the individual to report to the authorities rather than abet the concealment. This misalignment between "the corporation's interest in concealment" and "the individual's interest in immunity" is a structure analogous to the Prisoner's Dilemma, intended to function as an institutional guarantee for promoting information disclosure. Furthermore, the authorities retain the power to flexibly modify, suspend, or revoke authorization based on operational data. This represents a dynamic regulatory model that emphasizes performance during the operational phase rather than relying on a one-time pre-authorization.

California, operating under manufacturer self-certification, allows updates to be implemented at the developer's discretion as a matter of principle. The DMV reserves oversight authority during the operational phase as a condition of the deployment permit, seeking to ensure safety ex post through data disclosure obligations such as accident reports and disengagement reports. An example of this post-market monitoring model being actually enforced is the October 2023 suspension of the deployment permit for GM's Cruise.[35] When one of Cruise's driverless robotaxis was involved in an accident with a pedestrian, the company was criticized for failing to promptly disclose the full extent of the accident to the DMV—specifically, the fact that the vehicle had dragged the pedestrian. The DMV suspended the deployment permit, and the company halted driverless operations nationwide. This case demonstrates both that a post-market monitoring model based on data disclosure obligations can malfunction when disclosure is not properly conducted, and that the regulatory authority is capable of exercising the powerful tool of permit suspension. The UK's combination of the Duty of Candor and individual criminal liability is positioned as

[35] California DMV, Order of Suspension of Driverless Deployment Permit, Cruise LLC (October 24, 2023).

one institutional response to the challenges highlighted by the Cruise case—a design intended to deter information concealment ex ante.

3.2.2.3 Response to the Difficulty in Goal and Evaluation Methodology Setting

How to define "safety" is a fundamental challenge common to all three countries. The United Kingdom explicitly adopted the C&C standard (careful and competent driver) as a legal benchmark in the AVA, and Japan also references the C&C standard in its development of vehicle safety standards, drawing on discussions at the WP.29 FRAV (Framework Document on Automated/Autonomous Vehicles) (the SAKURA Project and the Ministry of Land, Infrastructure, Transport and Tourism's "Study Group on Safety Performance Assurance Measures for Autonomous Vehicles" final report (March 2025)).[36] The United States has not officially adopted such a unified qualitative standard, instead relying on a structure that leaves matters to developer self-assessment (VSSA). Nevertheless, all countries face the fundamental question of "by what criteria should the safety of AI be measured and judged," and common difficulties exist in translating this qualitative standard into an implementable form.

First, there is the "semantic gap" problem. The process of translating the abstract concept of "carefulness" demanded by law into technical specifications that engineers can program and verify is extremely challenging. Japan's Specified Autonomous Driving authorization system indirectly mitigates this gap by limiting operating conditions through ODD definitions, but the methodology for translating the C&C standard itself into technical specifications has only just begun. The United Kingdom, through the development of the Statement of Safety Principles (SoSP), is advancing efforts to decompose the C&C standard into more concrete safety principles, but this does not yet constitute a legally binding standard, and the final "translation" into technical specifications is left to developers.[37] In the United States, developers explain their own safety approach through the VSSA, but no unified standard exists for verifying its adequacy.

Second, there is the problem of legal predictability. As long as standards remain qualitative, the risk that courts may redefine safety standards based on "hindsight bias" after an accident occurs cannot be eliminated. This problem is directly connected to the question of how clearly the scope and limits of a safe harbor can be delineated.

---

[36] Ministry of Land, Infrastructure, Transport and Tourism, "Study Group on Safety Performance Assurance Measures for Autonomous Vehicles" Final Report (March 3, 2025). https://www.mlit.go.jp/jidosha/content/001879642.pdf

[37] Department for Transport (UK), Statement of Safety Principles for Automated Vehicles: Consultation Document (2025).

Third, there is the problem of safety evaluation methodologies themselves. While it is recognized that statistical accident rates alone cannot adequately evaluate safety, how to position alternative evaluation methods—such as formal methods and runtime monitoring—within the existing legal framework of the "duty of care" remains unestablished in any of these jurisdictions.

3.2.3 Summary

The regulatory system for autonomous driving is in the midst of a paradigm shift from "safety inspection of devices" to "governance of organizations, processes, and behavior." With respect to the black-box nature, all three countries have abandoned efforts to elucidate internal decision-making processes and instead employ indirect methods. Japan relies on DSSAD log recording and scenario-based evaluation, the United Kingdom on legal fiction attributing results to the ASDE, and the United States on driving performance data disclosure obligations. All rely on ex post verification of "behavior," with the UK ASDE model being notable for integrating this with organizational governance assessment.

With respect to dynamic mutability, Japan's snapshot assessment (version locking plus change notification) and the UK's dynamic management model (Duty of Candor plus individual criminal liability plus flexible modification, suspension, and revocation of authorization) present a stark contrast. The United States leaves matters to developer judgment and supplements with ex post data disclosure. In practice, Cruise had its operating permit revoked, partly due to inadequate information disclosure.

With respect to the difficulty in goal-setting, the United Kingdom has explicitly adopted the C&C standard and Japan also references it, while the United States lacks a unified standard. In all countries, the methodology for translating qualitative safety standards into verifiable technical specifications remains unestablished, and the "semantic gap" impedes the assurance of legal predictability.

3.3 Synthesis of the Case Studies: Institutional Implications Drawn from the Two Domains

This chapter has analyzed how the regulatory systems of major countries address the three technical challenges of UHAI---black-box nature, dynamic mutability, and difficulty in goal and evaluation methodology setting---focusing on the two domains of medical device AI (3.1) and autonomous driving (3.2). This section organizes these findings in the context of the three institutional challenges raised in Chapter 2: evaluation, certification, and legal frameworks.

3.3.1 Findings Concerning Evaluation

The regulatory systems in both domains have converged on safety evaluation through input-output verification---that is, "behavioral verification"---on the premise that directly requiring the elucidation of AI's internal structure is impossible. In the medical device domain, the post-lock evaluation under the IDATEN framework and the protocol review under the PCCP are concrete manifestations; in the autonomous driving domain, log-based evaluation through the DSSAD and scenario-based C&C standards serve the same function. This points to a common direction of addressing the black-box nature not through internal elucidation but through external verification.

However, this input-output evaluation methodology---systematic V&V (verification and validation) responsive to AI-specific risks---has not been established. As a complement to such product-level evaluation limitations, a common trend observed across both domains is the move to evaluate organizational management capabilities alongside product evaluation. The dual evaluation structure of the PCCP analyzed in Section 3.1 (appropriateness of the change protocol plus the organization's change management capability) and the two-tier structure of the UK ASDE analyzed in Section 3.2 (vehicle type approval plus organizational authorization) point to a shared direction of ensuring the safety of black-box products not solely through test results for the product alone but through a combination with the governance structure of the organization that develops and operates the product.

With respect to dynamic mutability, the challenge is how to overcome the limitations of one-time snapshot assessment. The TPLC approach (management throughout the total product lifecycle) under the US PCCP framework and the UK's Duty of Candor point to the direction of continuous evaluation throughout the entire lifecycle, though in both cases adoption in practice remains in its early stages. The experience of Germany's DiGA confirms that simplifying the entry point alone is insufficient and that the development of post-market monitoring infrastructure is indispensable to the credibility of the system.

With respect to the difficulty in goal-setting, the institutional mechanisms for forming social consensus on "what constitutes safe" remain unestablished. In the autonomous driving context, while the direction of a comparative benchmark---"equivalent to or better than a human expert"---is shared, neither the methodology for setting statistical thresholds nor the methods for measuring qualitative requirements such as "fairness" and "explainability" have been established.

These findings suggest the need for the standardization of V&V methodologies (the combined use of red teaming, metamorphic testing, formal methods, and other techniques. See 4.1.2 below), the institutionalization of continuous monitoring

throughout the entire lifecycle, and the formation of social consensus on acceptable risk levels.

3.3.2 Findings Concerning Certification

An important lesson that emerges from the analysis of both domains is that while product certification alone cannot continuously ensure the safety of dynamic systems such as AI, organizational certification alone also cannot guarantee product safety.

Current systems are moving toward the integration of product evaluation and organizational evaluation on the basis of this recognition. The US PCCP combines the review of product-specific risk management measures with the pre-assessment of the organization's change management capabilities. The UK ASDE model adopts a two-tier structure of vehicle type certification and organizational authorization. The approval structure of IDx-DR---the integration of the developer's responsibility with training requirements for the facility of use---is also a precedent in this direction.

At the same time, the core challenge in certification is the insufficiency of capability and methodology for regulatory authorities and certification bodies to independently evaluate the safety of black-box AI systems. In the medical device domain, the PMDA and the FDA possess the capacity to review clinical data, but their expertise in the technical evaluation of AI algorithms themselves is limited. In the autonomous driving domain, Japan's prefectural Public Safety Commissions face limitations in their capacity for technical evaluation of AI systems, and the US NHTSA relies on manufacturer self-certification. Even in the United Kingdom, the methodology for the VCA (Vehicle Certification Agency) to directly evaluate the safety of AI algorithms is still under development. In other words, while the direction of "integrated evaluation of product and organization" is becoming visible, the capacity-building of the entities that would conduct such evaluation has not kept pace.

Furthermore, the international frameworks for institutionally connecting domain-specific certification (ISO 13485 (Quality Management System standard for medical devices), vehicle type approval, etc.) with AI-specific certification (ISO/IEC 42001, etc.) are also still under development. The revision of ISO 13485 is not scheduled until around 2030, and institutional integration between the two is not expected in the near term.

These findings suggest the need for the institutional design of "joint certification" that integrates product risk management measures and organizational management capabilities in a unified evaluation, the building of technical capabilities at certification bodies, and the efficient connection with existing domain certification and the assurance of international interoperability.

3.3.3 Findings Concerning Legal Frameworks

The case studies of both domains yield the following implications for legal frameworks.

First, as a legal response to the black-box nature, countries are moving in the direction of abandoning the pursuit of causation within AI's internals. The legal fiction of the UK ASDE is the most explicit expression of this, circumventing the pursuit of causation by translating the technical black box into organizational responsibility. The possibility that "protocol compliance" under the PCCP may function as a safe harbor in the future and the significance of the De Novo approval of IDx-DR as a precedent for the legal acceptance of autonomous AI decision-making are both important materials for exploring the intersection of the black-box nature with authorization and liability regimes. Particularly noteworthy is the "compliance as defence" structure introduced by the UK AVA. A mechanism has been established whereby, if the ASDE has fulfilled authorization requirements, complied with the Duty of Candor, and taken reasonable safety management measures, it may be relieved of criminal liability as a due diligence defence. This demonstrates institutionally that, even where the internal causation of a black-box AI cannot be elucidated, the satisfaction of an organizational governance structure can itself function as a form of safe harbor, and it constitutes an important reference point for the design of the liability regime for UHAI.

Second, as a legal response to dynamic mutability, a recognition is becoming shared across countries that one-time pre-market assessment alone is insufficient and that continuous management throughout the entire product lifecycle is indispensable. However, it is neither technically nor resource-wise realistic for external regulatory authorities to directly and continuously monitor the state of AI systems that change on an ongoing basis. The direction toward which the systems of various countries are therefore converging is one of imposing the primary responsibility for continuous management on the operators themselves while establishing institutional guarantees to make that self-management effective. The UK's Duty of Candor mandates that operators self-report safety issues and imposes criminal liability on individual senior managers for false reporting or concealment, thereby structurally securing the incentive for self-reporting. The RWP monitoring obligation under the US PCCP similarly institutionalizes the operator's responsibility to continuously verify post-market changes on its own. In essence, the key to the legal response to dynamic mutability lies not in "comprehensive external supervision" but in ensuring effectiveness through the combination of "operator self-management plus reporting obligations plus sanctions for violations."

Third, as a legal response to the difficulty in goal-setting, the systems of various countries are moving toward constructing the consensus on "what constitutes safe" as a framework that integrates product performance, testing methodologies, and organizational structures. The "semantic gap" of qualitative safety standards,

exemplified by the C&C standard, is not merely a problem of difficulty in translating into technical specifications; it is directly connected to the problem of being unable to secure ex ante legal predictability---that is, the foreseeability of how far an operator must go to be relieved of legal liability. This indeterminacy also carries the risk of hindsight bias following an accident. In response to these challenges, the pre-authorization of "performance criteria plus verification methods plus acceptance criteria" under the PCCP, the elaboration of C&C standards through the UK's SoSP (Statement of Safety Principles), and the ASDE's due diligence defence can all be understood as attempts to delineate a safe harbor in an integrated manner across three dimensions: "what performance the product should demonstrate," "in what test environment it should be verified," and "what management system the organization should maintain." In the context of UHAI as well, the institutional efforts of various countries suggest that the design of safe harbors must be conducted not through any one of product, testing, or organization alone, but in an integrated form encompassing all three.

The institutional responses of various countries examined in this chapter present partial solutions for UHAI, but all remain under development. Accordingly, Chapter 4, drawing on these institutional innovations, envisions a framework in which evaluation criteria (4.1), certification mechanisms (4.2), and legal frameworks (4.3) for advancing the implementation of UHAI each function independently while also complementing one another.

[Reference Case] Anomaly Detection in Space Systems and AI Safety Evaluation

Although in a domain distinct from the medical device AI and autonomous driving examined in this chapter, the operation of space systems is also grappling with responses to the three fundamental challenges. In research on the thermal control system (TCA-L) of the International Space Station (ISS) Japanese Experiment Module "Kibo" (Iino et al.), a methodology for anomaly detection that considers the interdependencies among multiple parameters was developed. For a pump inverter malfunction caused by condensation that occurred in 2012, the approach detected signs of the anomaly in advance through increases in prediction error between a normal-behaviour model based on operational data and actual measurements, and further combined this with a formal method called SpecTRM-RL to identify the

cause of the anomaly from the characteristics of the data used to detect the precursory signs.[38]

What this research suggests is one practical approach to the three challenges. First, with respect to the black-box problem, by combining the systematic selection of critical parameters through FRAM (Functional Resonance Analysis Method) with ex post verification through SpecTRM-RL (a formal method), explainability for AI decisions is secured. Specifically, by using SpecTRM-RL to compare parameter condition combinations during normal and abnormal states, it was possible to identify condensation as the cause of the pump inverter failure, and this result was consistent with the troubleshooting actually conducted at the time. Second, with respect to dynamic mutability, an adaptive framework is adopted that uses a random forest normal-behaviour model to continuously monitor prediction errors and updates the FRAM model and adds parameters in response to environmental changes. Third, with respect to the setting of evaluation criteria, rather than alerts based on simple threshold exceedance, an approach is adopted that sets statistical thresholds based on a systems-theoretic understanding of the interdependencies among multiple parameters, and Pugh Concept Selection is used for model selection with multi-dimensional evaluation of detection accuracy, detection speed, and prediction accuracy.

"Kibo" orbits at approximately 400 km above the Earth and is continuously monitored around the clock from the Tsukuba Space Centre. Ground controllers must make complex judgments about conditions, including signs of anomalies, from hundreds of telemetry data points. The skill required depends heavily on each controller's expertise and experience, and the demand for machines and computers to support such complex tasks is growing amid current social conditions. This environment, in which immediate human intervention is constrained by communication delays and physical distance, can be positioned as a precursor to the "environments in which the premise of human involvement is difficult to maintain" that UHAI raises. The combination of "explainable AI monitoring plus verification through formal methods" demonstrated by this research suggests a technical direction for the concretization of process evaluation criteria, a common

[38] Iino, S., Nomoto, H., Fukui, T., Ishizawa, S., Sasaki, M., Yagisawa, Y., Hirose, T., Michiura, Y. & Shibayama, H. (2023). Systemic symptom detection in telemetry of ISS with explainability using FRAM and SpecTRM. Proceedings of 4th Asia Pacific Conference of the Prognostics and Health Management Society (PHMAP 2023), OS02-03; Iino, S., Nomoto, H., Fukui, T., Yagisawa, Y., Ishizawa, S., Hirose, T. & Michiura, Y. (2024). Towards Explainable Anomaly Detection in Safety-critical Systems Employing FRAM and SpecTRM in International Space Station Telemetry. International Journal of Prognostics and Health Management, 15(3). doi:10.36001/ijphm.2024.v15i3.3857.

challenge facing the regulatory systems of various countries examined in this chapter.

4. Key Issues for Future Institutional Design

Chapter 2 argued that the three technical characteristics of UHAI—black-box nature, dynamic mutability, and difficulty in goal and evaluation methodology setting—give rise to structural challenges across the three layers of evaluation, certification, and legal frameworks. The case studies in Chapter 3 confirmed that while the regulatory systems of various countries for medical device AI and autonomous driving partially address these challenges, all presuppose human involvement in some form and remain insufficient for UHAI.

Drawing on these analyses, this chapter organizes the issues that should be discussed going forward in the construction of institutions to enable the social implementation of UHAI, structured along the three layers of evaluation (4.1), certification (4.2), and legal frameworks (4.3). Section 4.4 then discusses how the three layers should interrelate and complement one another, and presents a perspective for breaking the "vicious cycle" identified in Chapter 2.

4.1 Issues Concerning the Evaluation of UHAI

4.1.1 Social Consensus on Acceptable Risk Levels

The most fundamental question in the evaluation of UHAI is: "What constitutes safe?" Given that AI is a system that operates probabilistically, guaranteeing zero risk is impossible in principle, and defining the standard of "what level of risk is socially acceptable" serves as the starting point for all evaluation.

As confirmed in Chapter 3, the UK's C&C standard (careful and competent driver) and Japan's SAKURA Project point to the direction of a comparative benchmark: "equivalent to or better than a human expert." This direction is a compelling option for UHAI, for which the setting of absolute safety standards is difficult. However, "equivalent to or better" here does not mean that every individual judgment must coincide with a human judgment. The possibility that AI may commit errors different from those of humans in individual judgments is not excluded; the condition for acceptance is that, when evaluated statistically as an aggregate of outcomes, safety equivalent to or better than that of a group of humans with appropriate competence is

ensured. That is, what should be assessed is not the identity of individual behaviors but the comparison of overall risk levels.

However, this comparative benchmark itself involves issues that require examination. First, the level of "humans with appropriate competence" varies substantially across domains, making it difficult to set a uniform baseline for comparison. Second, the methodology for statistically verifying whether "equivalent to or better" is achieved—including sample size, the setting of comparison conditions, and criteria for judging statistical significance—has not been established. Third, the methodology for measuring and evaluating qualitative requirements such as "fairness" and "explainability" in addition to physical safety remains undeveloped.

Furthermore, the determination of acceptable risk levels is not a purely technical judgment but involves ethical and political considerations. How much risk of what kind to accept is ultimately a matter of social choice. Accordingly, safety objectives need to be formulated through a highly transparent process involving diverse stakeholders, including not only government, business operators, and academia, but also consumer organizations and victim advocacy groups. How to institutionalize this consensus-building process—whether the public or private sector should lead the discussion, how it should be designed for each domain, and how to achieve international harmonization—is one of the most important issues in the construction of evaluation criteria.

4.1.2 Integration of Product Evaluation and Organizational Evaluation

The analysis in Chapter 3 demonstrated that product input-output verification ("behavioral verification") alone cannot sufficiently ensure the safety of black-box AI. This is because input-output verification is merely a result under specific test conditions, and it is impossible in principle to cover every situation that may be encountered during actual operation.

One direction for addressing this limitation is the advancement of product-level evaluation methodologies. Conventional safety testing for medical devices and automobiles verified whether the product "operates in accordance with specifications," but for UHAI, "specifications" themselves cannot be fixed, necessitating a redefinition of the subject of verification. Specifically, attention is being directed toward methodologies such as metamorphic testing (verifying whether input-output relationships remain stable under certain conditions), formal methods (mathematically proving behavioral properties, including neural network verification, to demonstrate the absence of deviation from the safe operating domain), and red teaming (discovering vulnerabilities by deliberately providing adversarial inputs).[39]

---

[39] On metamorphic testing, see Segura, S., Fraser, G., Sanchez, A.B. & Ruiz-Cortés, A. (2016). A Survey on Metamorphic Testing. IEEE Transactions on Software Engineering, 42(9), 805–824. For a survey of

The anomaly detection research on the ISS "Kibo" module introduced as a reference case in Chapter 3 demonstrates that by using formal methods (SpecTRM-RL), it is technically possible to ensure explainability for AI decisions even in environments where immediate human intervention is constrained, and this is suggestive for thinking about the direction of V&V methodologies for UHAI. However, how to incorporate these methods into the certification process, and what level of verification should be deemed "sufficient," remain as core questions in the standardization of V&V.

However, no matter how sophisticated product-level evaluation methodologies become, the fundamental limitation that they can only guarantee safety under test conditions is difficult to overcome. As a complement to this limitation, as confirmed in Section 3.3.1, the direction of evaluating "the management capabilities of the organization that develops and operates the product" alongside product evaluation is observed in common across US medical device regulation (the dual evaluation structure of the PCCP) and UK autonomous driving regulation (the two-tier structure of the ASDE).

The issue to be examined here is the question of how the relative weighting of product evaluation and organizational evaluation should be designed. If too much weight is placed on organizational evaluation, risks specific to individual products may be overlooked. On the other hand, if one insists on product input-output evaluation alone, one cannot escape the limitations of snapshot assessment and cannot respond to the dynamic nature of AI. The dual structure that the PCCP has arrived at—"appropriateness of the change protocol (product level) plus capability to execute the protocol (organizational level)"—points to one direction for balancing the two, but the optimal balance may differ depending on the nature of risks and the domain. Formulating criteria for when product evaluation should be emphasized and when organizational evaluation may be relied upon is a task for future examination.

4.1.3 Continuous Evaluation Throughout the Product Lifecycle

The limitations of snapshot assessment in the face of dynamic mutability were repeatedly confirmed in Chapter 3. Accordingly, the evaluation of UHAI must be designed as an integrated system of pre-market one-time assessment and post-market continuous monitoring. The issues can be organized into the following two categories.

The first is the institutionalization of post-market monitoring. As seen in the PCCP's RWP (Real-World Performance) monitoring and the UK's Duty of Candor, the direction of incorporating continuous post-market safety verification into the

---

formal methods for neural network verification, see Meng, W., et al. (2025). Adversarial Robustness of Deep Neural Networks: A Survey from a Formal Verification Perspective. IEEE Transactions on Dependable and Secure Computing, 22(1), 243–264.

institutional framework is clear, but the specifics of its design—the frequency of monitoring, indicators, thresholds, and the scope of reporting obligations—have not been established. Moreover, it is insufficient to leave monitoring solely to individual business operators; the development of public infrastructure for cross-industry sharing and analysis of accident and near-miss information should be considered. Coordination with international information-sharing frameworks such as the OECD's "Global AI Incident Reporting Framework"[40] is also an issue.

The second is the institutional connection between pre-market assessment and post-market monitoring. Under current systems, pre-market authorization and post-market monitoring are often institutionally separated. However, for UHAI, an important design challenge is how post-market monitoring results should be linked to the maintenance, modification, or revocation of authorization. How to connect runtime monitoring—which continuously monitors whether AI deviates from its safe operating domain during operation, and which transitions to safe fallback control upon detecting deviation (a "safety envelope" mechanism)[41]—with post-market monitoring and the authorization system is a core issue for the institutional design of dynamic authorization.

4.2 Issues Concerning the Certification of UHAI

4.2.1 Integration of Product Certification and Organizational Certification: The Potential of Joint Certification

As confirmed in Section 3.3.2, the failure of Pre-Cert (organizational certification alone is insufficient) and the formation of the PCCP's dual evaluation structure (product plus organization) empirically demonstrate that neither product certification nor organizational certification alone suffices, and that integration of the two is necessary.

One concept for institutionalizing this direction is "joint certification." Joint certification is a mechanism that evaluates as a package both the appropriateness of specific controls for the risks inherent in the AI system—such as bias countermeasures, testing methodologies, and safety envelopes—and the management capabilities of the organization to continuously operate and improve those controls—

---

[40] OECD, Towards a Common Reporting Framework for AI Incidents (2024). https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html

[41] On the concept of runtime monitoring / runtime verification and its application to autonomous systems, see Leucker, M. & Schallhart, C. (2009). A Brief Account of Runtime Verification. Journal of Logic and Algebraic Programming, 78(5), 293–303. On safe driving envelope verification, see Mehmed, A. (2021). Runtime Monitoring for Safe Automated Driving Systems. Doctoral Thesis, Chalmers University of Technology.

including quality management systems, change management processes, and incident response structures. In the context of international standards, the structure of ISO/IEC 42001 (AI Management Systems)—comprising the main text (organizational management requirements) and Annexes A/B (controls)[42]—could provide the foundation for such integrated evaluation.

Several issues arise regarding the concept of joint certification. First, in the review of product-specific risk management measures, how the adoption of V&V methodologies discussed in Section 4.1 (red teaming, formal methods, runtime monitoring, etc.) should be used as evaluation indicators. Second, that joint certification should be designed as a multi-tiered system calibrated to risk levels—rather than requiring identical rigorous certification for all AI, the depth and scope of certification required should be graduated according to the magnitude of the risk. Third, the potential for efficiency gains whereby, once organizational certification is obtained, only the product-specific risk management measures need undergo additional review for multiple AI systems provided by the same organization. This "reuse of organizational certification" is important from the perspectives of reducing certification costs and promoting innovation, but it needs to be examined in conjunction with the design of the validity period and renewal requirements of organizational certification.

Indeed, the seeds of this integrated approach are already visible in international standards and national frameworks. The EU AI Act mandates that providers of high-risk AI systems undergo conformity assessment for both product-level risk management and data governance (controls requirements) and quality management systems (organizational requirements).[43] The US NIST AI Risk Management Framework (AI RMF)[44] also adopts a structure in which organizational governance (Govern) serves as the foundation for practicing risk management of individual systems (Map, Measure, Manage). Within Japan as well—though relating to cloud government procurement requirements rather than AI—the "Information system Security Management and Assessment Program (ISMAP)" takes an approach that requires both organizational management structures and controls for individual systems, suggesting that the direction of joint certification is practically feasible.[45]

---

[42] ISO/IEC 42001:2023, Information technology — Artificial intelligence —Management system. This standard is the world's first international standard for AI management systems, comprising organizational management requirements (Clauses 4–10), AI-specific controls (Annex A), and implementation guidance (Annex B).

[43] Regulation (EU) 2024/1689, Articles 9, 17, 40–49.

[44] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0) (January 2023).

[45] National center of Incident readiness and Strategy for Cybersecurity (NISC), "Information system Security Management and Assessment Program (ISMAP)." https://www.ismap.go.jp/

4.2.2 Capacity-Building for Certification Bodies

An even more fundamental challenge for the institutional design of joint certification is the question of who can perform such certification. As confirmed in Section 3.3.2, the capacity of regulatory authorities and certification bodies to independently evaluate the safety of black-box AI systems is limited in every country.

In the medical device domain, the PMDA and the FDA possess the capacity to review clinical data, but their expertise in the technical evaluation of AI algorithms themselves—for example, bias assessment of training data, robustness verification of models, and prediction of performance degradation outside the training domain—is not sufficient. In the autonomous driving domain, Japan's prefectural Public Safety Commissions face structural limitations in their capacity for technical evaluation of AI systems, and the US NHTSA relies heavily on manufacturer self-certification. Even in the United Kingdom, the methodology for the VCA (Vehicle Certification Agency) to directly evaluate the safety of AI algorithms is still under development.

To bridge this capability gap, the following issues should be examined. First, the training and retention of personnel with AI technology expertise at certification bodies. Second, the institutionalization of international knowledge-sharing among certification bodies—frameworks that enable evaluation methodologies developed by one country's certification body to be utilized by bodies in other countries. Third, the possibility and conditions for leveraging private-sector specialist organizations in the certification process—building review structures through public-private partnerships. Even if the "form" of joint certification is designed, the system will not function if the entities implementing it lack the necessary capabilities. Capacity-building is an indispensable prerequisite for certification system design and should proceed in parallel with institutional design.

4.2.3 Integration with Existing Domain Certification and International Interoperability

UHAI will be deployed in existing regulatory domains such as medical devices, automobiles, and financial systems. Accordingly, rather than constructing AI-specific certification requirements from scratch, efficient integration with certification systems already existing in each domain is essential.

For example, ISO 13485, the quality management system standard for medical devices, and ISO/IEC 42001 for AI share the common structure of management systems based on PDCA cycles. If the system were designed so that a medical device manufacturer satisfying ISO 13485 could obtain joint certification by clearing only the additional requirements specific to AI (such as AI system impact assessment and data governance), efficient operation avoiding duplicative review could be achieved. However, realizing this requires revising existing standards or creating new ones, and institutional integration with AI-specific standards is not expected to be achieved in the short term

Given this situation, in the short to medium term, a pragmatic approach is needed that designs "interface connections" between each domain's certification system and AI certification without waiting for the integration of international standards. International standardization of conformity assessment schemes for AI systems is being advanced through ISO/IEC 42007[46] and other instruments, and it is important for Japan's certification system to maintain alignment with these in order to ensure global interoperability.

4.3 Issues Concerning the Legal Frameworks for UHAI

The analysis in Chapter 3 (Section 3.3.3) showed that the legal frameworks of various countries are commonly moving toward the following directions: abandoning the pursuit of causation within the black box and translating it into organizational responsibility; responding to dynamic mutability through a combination of operator self-management and sanctions; and constructing safe harbors that integrate product, testing, and organizational dimensions. Drawing on these findings, this section organizes the issues that should be examined in the legal framework design for UHAI, structured along the three dimensions of pre-market regulation, enforcement, and liability regime as organized in Chapter 2.

4.3.1 Pre-market Regulation: Legal Acceptance of the Black Box and Dynamic Authorization

Current authorization systems are premised on the applicant articulating the specifications of the system and the regulatory authority being able to examine their content. The three technical characteristics of UHAI—black-box nature, dynamic mutability, and difficulty in goal-setting—all undermine this premise, and the following issues arise regarding the redesign of pre-market regulation.

The first issue concerns the legal acceptance of the black box. If the V&V technologies discussed in Section 4.1.3 (such as neural network verification using formal methods) mature, can the technical assurance of safe operating domains be accepted as an authorization requirement in lieu of articulation of specifications? This question compels a reexamination of the very premises of the authorization system.

The second issue concerns the possibility of dynamic authorization. The PCCP analyzed in Chapter 3 is a system that permits "changes within a pre-planned scope," but the changes of UHAI may be neither planned nor scopeable. Whether a dynamic authorization framework—one that, by linking the runtime monitoring discussed in Section 4.1.3 with the authorization system, "continues authorization as long as

[46] ISO/IEC 42007 (under development), Information technology—Artificial intelligence—Application management system requirements for organizations providing or using AI.

changes remain within the safe domain and immediately halts operation if the domain is breached"—is legally viable is worth examining. Such a framework would transform authorization from a "one-time state review" to "continuous process approval," requiring institutional development.

The third issue concerns the relationship between goal-setting and authorization criteria. If the acceptable risk levels discussed in Section 4.1.1 are agreed upon socially, it is institutionally possible to make the satisfaction of those criteria a condition of authorization, but as long as the criteria are qualitative, authorization decisions will inevitably involve discretion. The UK's SoSP (Statement of Safety Principles) is an attempt to decompose qualitative criteria into more concrete safety principles, but the final translation into technical specifications is left to developers. How far this "semantic gap" can be closed institutionally—and from where it must inevitably be left to the developer's judgment—is an issue directly connected to the effectiveness of pre-market regulation.

4.3.2 Enforcement: Institutional Design of Self-management and Third-party Oversight

As noted in Chapter 2, enforcement against UHAI faces the triple challenge of monitoring difficulties due to the black-box nature, difficulty in identifying the moment of violation due to dynamic mutability, and the absence of enforcement standards due to the difficulty in goal-setting. The analysis in Chapter 3 showed that countries are moving not toward "comprehensive external supervision" but toward ensuring effectiveness through the combination of "operator self-management plus reporting obligations plus sanctions for violations."

When applying this direction to UHAI, the following issues should be examined.

The first is the institutional guarantee for making self-management effective. The UK's Duty of Candor analyzed in Chapter 3 mandates that operators self-report safety issues and imposes criminal liability on individual senior managers for false reporting or concealment, thereby creating a structural tension between the corporation and the individual that is analogous to a Prisoner's Dilemma. This institutional design is one model for ensuring the effectiveness of information disclosure in situations where reliance on operator self-management is unavoidable. However, for UHAI, there is an additional challenge: the operator itself may not be able to fully comprehend the system's changes.

The second is the complementary role of external oversight. Whether the institutionalization of continuous monitoring by independent third parties—for example, establishing an AI safety specialist agency comparable to an aviation accident investigation board—is necessary when self-management by operators alone has limitations is an issue directly connected to the design of the enforcement architecture. Such an agency would require the capability for technical evaluation of AI, the capability for accident cause investigation, and the authority to make

recommendations to regulatory agencies, but the cost-benefit balance of its establishment and operation must also be considered.

The third is the development of infrastructure for ex post verification. While the mandatory installation of DSSADs (Data Storage Systems for Automated Driving) in autonomous driving represents an attempt to secure records for ex post verification, it remains uncertain for UHAI how far the reconstruction of decision-making processes is possible from input-output logs alone. How to set recording requirements that enable "sufficient ex post verification" while abandoning "complete reconstruction" is an institutional design challenge relevant to both enforcement and the liability regime discussed in the following section. Even where data recording obligations exist, the post-market monitoring model will malfunction if disclosure is not properly conducted. The institutional connection between recording obligations and disclosure obligations is also an important issue.

4.3.3 Liability Regime: Integrated Design of Safe Harbors and New Categories of Liability

How legal liability should be allocated when UHAI causes harm is, together with evaluation and certification, a critical institutional condition that determines whether UHAI can be socially implemented. As analyzed in Chapter 2, the difficulty of proving causation due to the black-box nature, the inconsistency between dynamic mutability and the concept of "defect at the time of delivery," and the indeterminacy of the basis for negligence determination arising from the difficulty in goal-setting can all give rise to malfunctions in the liability regime. The analysis in Chapter 3 showed that countries are addressing these challenges through the directions of translating liability to organizational responsibility, mandating lifecycle management, and constructing safe harbors that integrate product, testing, and organizational dimensions. The following organizes the issues that should be examined in designing the UHAI liability regime.

The first issue concerns the integrated design of safe harbors. As confirmed in Chapter 3, the UK AVA's due diligence defence is a mechanism whereby the ASDE may be relieved of criminal liability if it has fulfilled authorization requirements, complied with the Duty of Candor, and taken reasonable safety management measures—an institutionalization of the "compliance as defence" structure. The pre-authorization of "performance criteria plus verification methods plus acceptance criteria" under the PCCP and the UK SoSP's elaboration of C&C standards can also be understood as attempts to delineate the contours of a safe harbor. These findings suggest that safe harbors for UHAI should be designed in an integrated form encompassing three dimensions: "what performance the product should demonstrate," "in what test environment it should be verified," and "what management system the organization should maintain." However, if the boundaries of the safe harbor remain unclear, the problem of hindsight bias—the risk that courts will redefine safety standards ex post after an accident—persists. How to clearly delineate the scope and limits of the safe

harbor and thereby secure legal predictability is a central issue in liability regime design.

The second issue concerns addressing the difficulty of proving causation. None of the case studies in Chapter 3 presented a method for directly elucidating the chain of causation within the black box. The UK ASDE's legal fiction presented one answer—circumventing the pursuit of causation altogether and attributing outcomes to the organization—but this relies on the existence of an organizational responsible entity as the predicate for the fiction, and is inseparable from the design of the responsible entity. The question to be examined here is whether procedural measures such as presumptions of causation or burden-shifting are sufficient, or whether a transition to categories of liability that do not require proof of causation—strict liability or no-fault liability[47]—should be considered. The extent to which the product liability framework can address the causation problems of UHAI is an issue that warrants close attention to future developments.

The third issue concerns the reconstruction of the "defect at the time of delivery" concept. The "defect at the time of delivery" requirement under the Product Liability Act (*Seizōbutsu Sekinin Hō*) creates a fundamental inconsistency with dynamically changing UHAI. If performance degrades or unexpected decision-making tendencies are acquired through post-delivery learning or environmental changes, current law provides no clear answer regarding "at what point in time the state should serve as the baseline for judging a defect." The continuous safety obligation under the UK ASDE model suggests the direction of extending the baseline for liability from "the time of delivery" to the entire lifecycle. If this direction is pursued, the question of how to design the allocation of responsibility among developers, providers, and operators at each stage of the lifecycle becomes a significant issue entailing the reconstruction of the liability regime across the entire supply chain.

4.4 Conclusion: Toward a Society That Maximizes the Benefits of UHAI

The foregoing has organized issues to be discussed going forward, rather than providing definitive answers. However, through the analysis of this report, directions for institutional design are becoming visible for each of the three layers of evaluation, certification, and legal frameworks. The following provides a concise summary as a conclusion to this report.

---

[47] On the possibility of introducing strict liability for AI, see Vladeck, D.C. (2014). Machines Without Principals: Liability Rules and Artificial Intelligence. Washington Law Review, 89(1), 117–150. For an overview of the EU's discussion on AI liability, see European Parliament (2020). Civil Liability Regime for Artificial Intelligence: European Added Value Assessment, PE 654.178.

First, regarding evaluation, the direction is to construct a new evaluation paradigm that responds to UHAI's three technical characteristics. To address the black-box nature, rather than attempting to elucidate AI's internals, the approach is to leverage new V&V methods such as formal methods and red teaming to standardize techniques for externally verifying AI "behavior." To address dynamic mutability, the approach is to design pre-market verification and post-market real-world performance monitoring as an integrated system, establishing mechanisms for continuous evaluation throughout the entire product lifecycle. To address the difficulty of goal-setting, the approach is not to pursue "absolute safety" but rather, anchored in a relative and statistical comparison with conventional human-operated processes, to formulate acceptable risk levels through a transparent process involving diverse stakeholders and to take that consensus as the starting point for evaluation criteria.

Second, regarding certification, the direction is to construct a certification system that effectively underpins the evaluation criteria described above. To address the black-box nature, a "joint certification" framework is to be constructed that examines not only the product's input-output verification but also the governance structure of the organization that develops and operates it. To address dynamic mutability, rather than requiring a fresh review from scratch with each change, the approach of pre-approving the change management process itself is to be advanced. To address the difficulty of goal-setting, socially agreed standards are to be connected to certification judgment criteria, and the capacity for certification bodies to independently verify compliance with those standards is to be built.

Third, regarding legal frameworks, the direction is to give legal effect to the outcomes of evaluation and certification and to secure legal predictability for business operators. To address the black-box nature, the technical assurance of safe operating domains is to be accepted as an authorization requirement in lieu of explanations of internal structure, and the maintenance of appropriate organizational structures is to be legally positioned as a safe harbor. To address dynamic mutability, operators are to be mandated to conduct continuous safety management and self-reporting, with the effectiveness of these obligations ensured through appropriate incentive structures for information provision. To address the difficulty of goal-setting, test environments and certification systems are to be developed that enable operators to demonstrate their achievement of socially agreed standards.

What is common to these three directions is that the three layers of evaluation, certification, and legal frameworks must not be designed in isolation but must be designed in an integrated manner so that they function as a unified whole. That is, consensus on evaluation criteria provides the foundation for certification; certification attainment provides the basis for safe harbors; and the clarification of safe harbors enhances operators' legal predictability and promotes innovation.

UHAI holds the potential to transcend the limits of human capability and deliver greater safety and benefits to society. To make this potential a reality, we must move beyond the conventional assumption that "safety is ensured by human oversight" and envision new institutional frameworks that can adapt to the evolution of technology. It is our hope that the issues organized in this report will serve as a starting point for that discussion.

Table: Technical Characteristics of UHAI (columns) and Corresponding Directions for Institutional Design (rows)

	Black-box nature	**Dynamic mutability**	**Difficulty in goal-setting**
Evaluation	Abandon elucidation of AI internals; **standardize techniques for externally verifying AI "behavior"** through new V&V methods such as formal methods and red teaming	**Design pre-market verification and post-market Real-World Performance (RWP) monitoring as an integrated system**, establishing continuous evaluation throughout the entire lifecycle	Rather than pursuing "absolute safety," build consensus on acceptable risk levels through a transparent process with diverse stakeholders, anchored in **relative and statistical comparison** with conventional human operations
Certification	Construct a **"joint certification" framework that reviews both** product input-output verification and organizational governance as a package	Rather than re-reviewing from scratch for each change, advance an approach that **pre-approves the change management process itself**	**Connect socially agreed standards to certification judgment criteria** and build the capacity for certification bodies to independently verify them
Legal frameworks	Accept **technical assurance of safe operating domains as an authorization requirement** in lieu of explaining internal structure; legally position maintenance of appropriate organizational structures as a **safe harbor**	**Mandate continuous safety management and self-reporting** by operators, ensuring effectiveness through appropriate incentivization of information provision	Develop **test environments and certification systems** that enable operators to demonstrate achievement of agreed standards, securing safe harbors that provide legal predictability

FY 2025 Study Group on Issues and Policies in the Application of AI
List of Committee Members
(in alphabetical order; honorifics omitted)

※Authors of the Report

■ **Chair**

- **Hiroki Habuka*** Research Professor
  The Center for Interdisciplinary Studies of Law and Policy
  Graduate School of Law, Kyoto University
  Attorney-at-Law

■ **Members**

- **Koichi Ito** PricewaterhouseCoopers Japan LLC Partner
  Chief of AI Audit Lab
  Certified Public Accountant (Japan)
- **Tatsuhiko Inatani** Professor, Graduate School of Law, Kyoto University
- **Takafumi Ochiai** Head of Policy Research Institute
  Senior Partner, Atsumi & Sakai
- **Roy Sugimura** Chief Coordination Officer, Coordination Office
  Department of Information and Human Factors
  National Institute of Advanced Science and Technology (AIST)
- **Kumiko Takahashi** Senior Researcher, Urban Infrastructure DX Group
  Social Infrastructure Division
  Mitsubishi Research Institute, Inc.
- **Kuan-Wei Chen*** Program-Specific Assistant Professor
  Graduate School of Law
  Kyoto University
- **Yuchang Cheng*** Senior Research Manager
  Data & Security Research Laboratory
  Fujitsu Limited
- **Keisuke Tomiyasu** CTO, AI Medical Service Inc.
- **Takayuki Hirose** Program-Specific Junior Associate Professor
  Kyoto University Graduate School of Law

■ **Observers**

- **Hiroki Takamura** Chief Researcher
  Information-technology Promotion Agency, Japan
  Digital Infrastructure Center
  Digital Engineering Department
  AI System Group

- **David Socol de la Osa David Uriel**
  Assistant Professor
  Hitotsubashi Institute for Advanced Study
  Hitotsubashi University

■ **Secretariat** Digital Society Research Institute
Center for International Economic Collaboration (CFIEC)

- **Masanobu Katoh** Head of the Institute
- **Eiichi Matsuzawa** Executive Expert
- **Morimasa Katagiri** Chief Research Fellow