Provisional translation for reference purposes only

Guideline for Responsible AI Application in Research and Development of AI-Equipped Defense Systems

(Ver. 1)

Ministry of Defense

Table of Contents

1 Background to the Formulation of the Guideline

The rapid development and widespread adoption of artificial intelligence (AI) technology in recent years have significantly improved AI performance, but new challenges have also emerged. AI operates not by clear rules or conditions set by humans, but by learning rules and knowledge from given data and outputting new results based on that. Because of this, AI has unique technical risks, such as the possibility of bias or misjudgment arising from biased training data. Furthermore, as AI performance has improved, the impact of AI on humans and society has grown. Beyond technical risks, there are ethical, legal, and social risks unique to AI, such as the opacity of AI's inference processes and judgment criteria, concerns about personal information protection and privacy, the possibility of copyright infringement, the generation and dissemination of fake information by AI, and the impact of AI adoption on employment. To address these issues, international discussions on the responsible research and development and use of AI technology are actively underway. Various governments, international organizations, companies, and academic bodies are working to formulate AI ethics guidelines and regulatory frameworks. The Ministry of Defense formulated the *MOD Artificial Intelligence Strategy* (Ministry of Defense, July 2024) (hereinafter referred to as the "Strategy") to convey the Ministry of Defense and Self-Defense Forces' approach to the use of AI both internally and externally. The Strategy stipulates that the Ministry of Defense and Self-Defense Forces will formulate our own guidelines regarding research and development.

(1) Initiatives of Various Countries Regarding the Responsible Use of AI in the Military Domain

Various countries are undertaking the following initiatives regarding the responsible use of AI in the military domain:

The U.S. Department of Defense formulated the *Ethical Principles for Artificial Intelligence* in 2020, advocating five principles: responsible, equitable, traceable, reliable, and governable. Based on these principles, it aims to ensure ethical considerations and technical reliability in the military use of AI. In 2022, the U.S. Department of Defense formulated the *Responsible Artificial Intelligence Strategy and Implementation Pathway*, which serves as a course of action for introducing responsible AI with the desired end state of achieving AI reliability. Subsequently, in 2023, the U.S. Department of Defense led the *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* (hereinafter referred to as the "Political Declaration on AI"), which was supported and joined by many countries.

In France, the Military Ethics Committee, established by the French Ministry of Armed Forces, compiled the *Opinion on the Integration of Autonomy into Lethal Weapons Systems* in 2021. The French Ministry of Armed Forces emphasizes the importance of human judgment in the military use of AI and opposes the research and development of fully autonomous lethal weapon systems. It also supports and participates in the U.S.-led Political Declaration on AI and is actively involved in forming international norms.

The United Kingdom formulated the *Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-enabled Capability in Defence* in 2022 and is advancing its review from the ethical, legal, and technical perspectives. The UK Ministry of Defence emphasizes the importance of human decision-making in the military use of AI and opposes the research and development of fully autonomous lethal weapon systems. The UK also supports and participates in the Political Declaration on AI and is actively involved in forming international norms.

(2) International Discussions on the Military Use of AI and Japan's Views

A. Overview of the Political Declaration on AI etc.

The Political Declaration on AI, launched under the leadership of the United States, is an important initiative in providing international guidance for ensuring the responsible development, deployment, and use of AI in the military domain. Participating states are expected to implement appropriate measures at each relevant stage throughout the entire lifecycle of military AI capabilities. Many countries, including Japan, have expressed support for the declaration. As of December 2024, 58 states have endorsed it.

Although the declaration is not legally binding, endorsing states are required to implement the measures shown in Table 1, engage in continuous consultations, and improve these measures to promote the objectives of the declaration regarding the military use of AI.

In addition, there is the *Responsible Artificial Intelligence in the Military Domain (REAIM)* initiative led by the Netherlands and the Republic of Korea, and the outcome document, supported by many countries including Japan, confirms that AI must be utilized responsibly in the military domain while complying with applicable international law.

B. Discussions on Lethal Autonomous Weapon Systems (CCW etc.)

Against the backdrop of rapid technological advancements in the military field in recent years and growing international interest, discussions on Lethal Autonomous Weapon Systems (LAWS) began in 2014 under the Convention on Certain Conventional Weapons (CCW) (hereinafter referred to as "CCW"). Since 2017, the Group of Governmental Experts (GGE) (hereinafter referred to as "GGE") established under the CCW has been leading discussions regarding the creation of international rules concerning LAWS.

Discussions at the CCW LAWS GGE mainly cover topics such as the characteristics and definition of LAWS, the applicability of international humanitarian law, the nature of human involvement, responsibility and accountability, risk mitigation and confidence building, and the nature of regulation. One of the key achievements is the adoption of the Guiding Principles adopted by the GGE in November 2019, consisting of 11 points, including that international humanitarian law applies to LAWS and that human responsibility must be retained. In May 2023, the LAWS GGE report was adopted, which describes considerations for prohibitions and restrictions from the perspective of compliance with international humanitarian law. However, an international definition of LAWS has not yet been established.

In December 2023, an Austrian draft resolution calling for the UN Secretary-General to prepare a report on LAWS was adopted by the UN General Assembly, and the UN requested opinions from UN member states. In response, in May 2024, Japan also submitted a working paper summarizing Japan's views (hereinafter referred to as the "*LAWS Working Paper*") to contribute to the preparation of the report and discussions at the CCW. In the *LAWS Working Paper*, Japan states that it will not conduct research and development or operations of weapon systems whose use is not permitted under international law, including international humanitarian law, and domestic law. It also presents Japan's views on weapon systems whose development and use should not be permitted internationally.

Table 1: Measures to be Implemented by Endorsing States in the Political Declaration on AI

A	States should ensure their military organizations adopt and implement these principles for the responsible development, deployment, and use of AI capabilities.
B	States should take appropriate steps, such as legal reviews, to ensure that their military AI capabilities will be used consistent with their respective obligations under international law, in particular international humanitarian law. States should also consider how to use military AI capabilities to enhance their implementation of international humanitarian law and to improve the protection of civilians and civilian objects in armed conflicts.
C	States should ensure that senior officials effectively and appropriately oversee the development and deployment of military AI capabilities with high-consequence applications, including, but not limited to, such weapon systems.
D	States should take proactive steps to minimize unintended bias in military AI capabilities.
E	States should ensure that relevant personnel exercise appropriate care in the development, deployment, and use of military AI capabilities, including weapon systems incorporating such capabilities.
F	States should ensure that military AI capabilities are developed with methodologies, data sources, design procedures, and documentation that are transparent to and auditable by the relevant defense personnel.
G	States should ensure that personnel who use or approve the use of military AI capabilities are trained so they sufficiently understand the capabilities and limitations of those systems in order to make appropriate context-informed judgments on the use of those systems and to mitigate the risk of automation bias.
H	States should ensure that military AI capabilities have explicit, well-defined uses and that they are designed and engineered to fulfill those intended functions.
I	States should ensure that the safety, security, and effectiveness of military AI capabilities are subject to appropriate and rigorous testing and assurance within their well-defined uses and across their entire life-cycles. For self-learning or continuously updating military AI capabilities, States should ensure that critical safety features have not been degraded, through processes such as monitoring.

Table 1: Measures to be Implemented by Endorsing States in the Political Declaration on AI

(Cont.)

J	States should implement appropriate safeguards to mitigate risks of failures in military AI capabilities, such as the ability to detect and avoid unintended consequences and the ability to respond, for example by disengaging or deactivating deployed systems, when such systems demonstrate unintended behavior.

2 Significance of the Guideline

With concrete examples of AI application to equipment increasing in a variety of countries, such as the deployment of AI-equipped unmanned aerial vehicles in actual combat and the use of AI for information gathering and analysis, the need for effective rulemaking for responsible AI application in the defense sector is growing. Furthermore, while weapon systems with autonomy that ensures human involvement can provide significant benefits for security, such as reducing human error and labor, AI presents unique difficulties distinct from conventional technologies, such as the opacity of judgment processes, the possibility of unexpected behavior, and ethical issues. While methods for responsible AI application in the military and defense fields are being carefully considered in a variety of countries as introduced in the previous section, frameworks that specifically indicate measures to be implemented, such as the Political Declaration on AI of the U.S., are emerging. In response to these developments, the Ministry of Defense's Strategy clarified the policy of formulating guidelines for research and development of equipment.

Against this backdrop, this *Guideline for Responsible AI Application in Research and Development of AI-Equipped Defense Systems* (hereinafter referred to as "the Guideline") aims to materialize a series of commitments of the Ministry of Defense and Self-Defense Forces and provide a framework for appropriately managing and mitigating AI risks in the context of research and development. This is intended to ensure predictability for businesses participating in research and development and further promote the utilization of AI in defense equipment. The Guideline covers research and development projects from the conceptual stage to the research and development stage within the equipment lifecycle shown in Figure 1. The definition of AI in the Guideline adheres to the Strategy and refers to machine learning software and programs.

It should be noted that the Guideline is regarded as a guideline that implementers should comply with when planning and implementing research and development projects related to defense equipment to which AI technology is applied (hereinafter referred to as "AI-equipped systems").

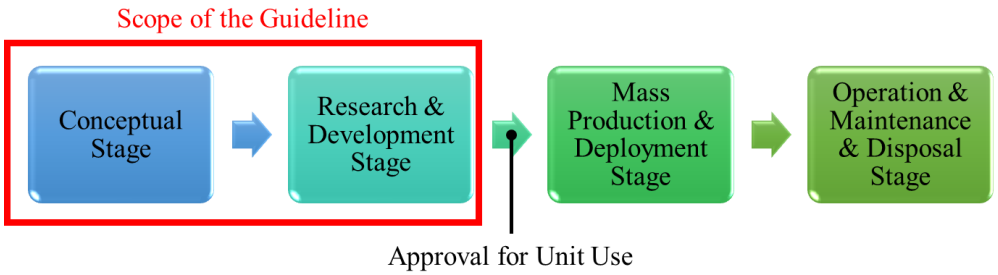

Figure 1: Scope of the Guideline

3 Confirmation Points in Research and Development of AI-Equipped Systems

Based on the *LAWS Working Paper* mentioned in Section 1, to enable management that ensures effective measures against AI application risks (hereinafter referred to as "risk management"), the requirements to be complied within the research and development of AI-equipped systems are set as follows and whether these requirements are fulfilled will be confirmed during the research and development project.

(1) Establishment of Applicable Requirements

Japan considers that the principles of international humanitarian law apply to all weapons, including those utilizing emerging technologies, such as AI. Furthermore, as a matter of course, the Ministry of Defense and Self-Defense Forces will not conduct research and development or introduce equipment whose use is not permitted under international law or domestic law. In the discussions on LAWS, Japan has also stated that weapon systems which cannot be used in accordance with international humanitarian law or autonomous weapon systems with lethal force that operate completely without human involvement have the potential to cause unacceptable consequences; therefore, their development and use should not be permitted internationally. Therefore, the operational concept of the system under research and development should comply with this.

Based on this, the operational concept of AI-equipped systems will be checked to see whether it is appropriate from legal/policy perspectives (hereinafter referred to as "Requirement A") and whether it satisfies the operational concept from a technical perspective (hereinafter referred to as "Requirement B"). A-1 is a requirement from a legal perspective on whether it satisfies the requirements of international law, including international humanitarian law, and domestic law. Although Requirement A-1 is not a requirement specific to the research and development of AI-equipped systems, given Japan's position that the principles of international humanitarian law apply to all weapons, including those utilizing emerging technologies, and that it will not conduct research and development of equipment whose use is not permitted under international or domestic law, it is listed as one of the requirements to be confirmed in this document because compliance with international law including international humanitarian law has been a focus of discussions on weapons utilizing emerging technologies. A-2 is a requirement that reflects Japan's policy stance, keeping international humanitarian law in mind. It should be noted that A-2 is a requirement reflecting Japan's current policy stance, and its revision will be considered as necessary if new ideas that Japan believes should be relied upon are formed in the course of future discussions in international forums.

Furthermore, in discussions concerning responsible AI application, Requirement B, consisting of the elements B-1 to B-7 shown below, should also be complied with. Requirement B is introduced to confirm, from a technical perspective, whether prototypes of high-risk AI-equipped systems (described later) that are deemed to satisfy Requirement A at the conceptual stage are equipped with functions that enable them to continue to satisfy Requirement A even at the actual deployment and operation stages and whether appropriate risk mitigation measures have been implemented. In the United States and other countries, the desired direction for responsible AI implementation is articulated in the form of AI ethics principles. As shown in Figure 2, Requirement B is consistent with international trends in its relationship with AI ethics principles in a variety of countries. For risk management in research and development projects, whether Requirements A and B are fulfilled will be continuously confirmed at necessary milestones, such as reviews, from the conceptual stage to the completion of research and development.

Legal/Policy Requirements (Requirement A)

A-1 It is not a project that is non-compliant with international law, including international humanitarian law, and domestic law.

Not something that causes superfluous injury or unnecessary suffering by its nature, is inherently indiscriminate, or cannot be used in accordance with international humanitarian law.

Something that can be used in accordance with domestic law.

A-2 It is not an autonomous weapon with lethal force that operates completely without human involvement.

Not something that is autonomous with lethal force that operates completely without human involvement, which does not involve an appropriate level of human judgment and which cannot be operated within a responsible chain of human command and control.

Technical Requirements (Requirement B)

B-1 Clarification of Human Responsibility

Designed to allow human operators to exercise appropriate levels of care and act responsibly when using AI systems, enabling operator involvement and appropriate control by the operator.

B-2 Fostering Operator's Appropriate Understanding

Designed with mechanisms to enable operators to use AI systems appropriately, measures to prevent excessive reliance, and mechanisms for operators to improve the system in the case of malfunctions.

B-3 Ensuring Fairness

Appropriate mitigation measures are implemented for AI systems and datasets after exploring and understanding the causes of bias, thereby minimizing undue and harmful bias.

B-4 Ensuring Verifiability and Transparency

The system construction process, such as the design procedures, adopted algorithms, and data used for learning, is clarified, ensuring the verifiability and transparency of AI systems.

B-5 Ensuring Reliability and Validity

AI system reliability, effectiveness, and security are evaluated through testing from various perspectives, ensuring operations at acceptable levels throughout the entire lifecycle.

B-6 Ensuring Safety

Mechanisms are in place to reduce the risk of AI system malfunction or serious failure, thereby safety is ensured.

B-7 Compliance with International Law and Domestic Law

Designed to enable operation in compliance with applicable international law and domestic law.

		Concepts Cited in Policy Documents of Various Countries			
		U.S. (AI Principles) (Feb. 2020)	Australia (A Method for Ethical AI in Defence) (Jan. 2021)	UK (Ambitious, Safe, Responsible) (Jun. 2022)	France (Opinion on the Integration of Autonomy into Lethal Weapon Systems) (Apr. 2021)
B-1	Clarification of Human Responsibility	Responsible	Responsibility	Responsibility	Command
B-2	Fostering Operator's Appropriate Understanding	—		Understanding	Competence
B-3	Ensuring Fairness	Equitable	—	Bias and Harm Mitigation	—
B-4	Ensuring Verifiability and Transparency	Traceable	Traceability	—	—
B-5	Ensuring Reliability and Validity	Reliable	Trust	Reliability	Confidence
B-6	Ensuring Safety	Governable	Governance	—	—
B-7	Compliance with International Law and Domestic Law	—	Law	Human-Centric	Compliance

Figure 2: AI Ethics Principles in the Military Domains of Various Countries

4 Measures to Implement in Research and Development of AI-Equipped Systems

The following section outlines the process for confirming whether the requirements set in the previous section are fulfilled. Specifically, the three points shown in Figure 3, "Classification of AI-Equipped Systems," "Legal/Policy Review," and "Technical Review," will be implemented. Details of the review procedures and institutional frameworks will be stipulated separately.

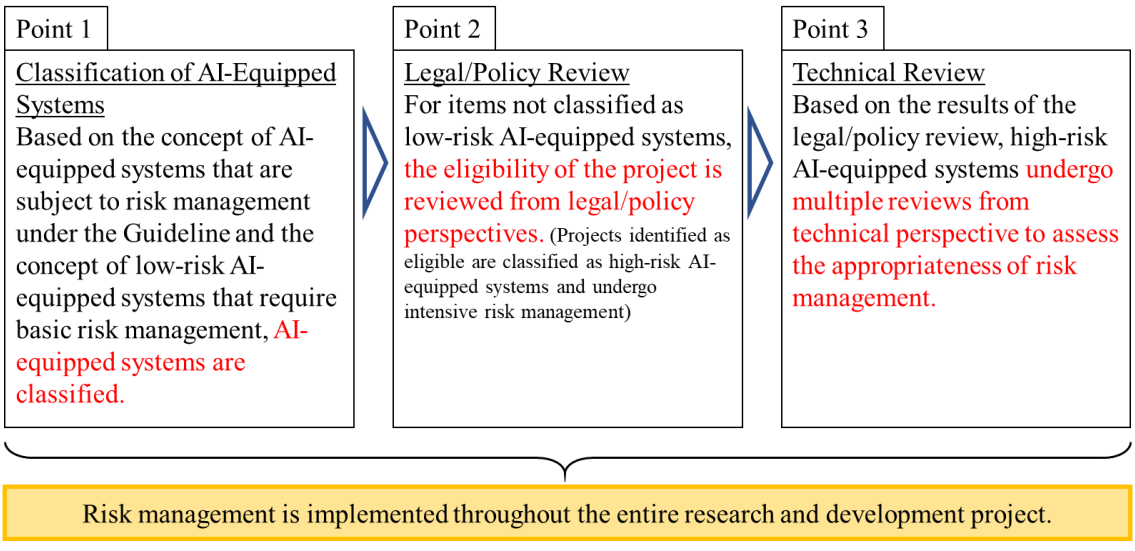

Figure 3: Implementation Measures in Research and Development Projects

(1) Classification of AI-Equipped Systems

In the research and development of AI-equipped systems, equipment that particularly requires intensive risk management from the perspective of the LAWS mentioned in Section 1 will be classified as high-risk AI-equipped systems, while other AI-equipped systems will be classified as low-risk AI-equipped systems, and appropriate risk management will be implemented according to the type of equipment.

Specifically, equipment will be classified according to the flow in Figure 4. While there are research and development projects for a variety of different types of equipment, the Guideline primarily targets prototype projects for equipment using AI technology. Furthermore, if the output resulting from the AI function of the equipment does not affect functions possessing destructive capability within the system in which the equipment is included, it will be classified as a low-risk AI-equipped system. Otherwise, it will be subject to a legal/policy review (details described later), and if identified as ineligible, the research and development of that equipment will be canceled at that point because it may cause unacceptable consequences. If identified as eligible, it will be classified as a high-risk AI-equipped system, and a technical review (details described later) will be conducted to implement intensive risk management.

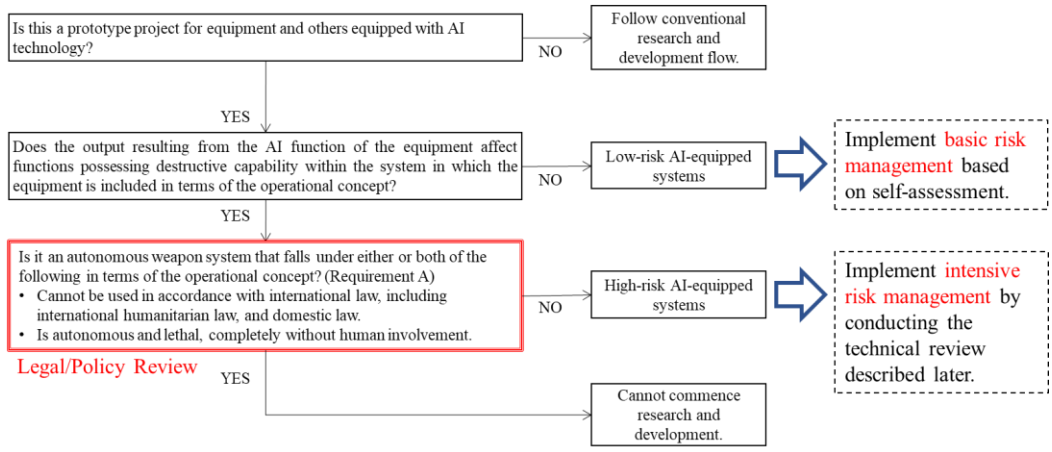

Figure 4: Equipment Classification and Response Flow

(2) Legal/Policy Review

The legal/policy review is conducted at the conceptual stage of AI-equipped systems to assess whether the planned functions and envisioned modes of use of the AI-equipped system under research and development are appropriate in accordance with Requirement A mentioned in the previous section from legal/policy perspectives. The review targets AI-equipped systems that were not judged as low-risk AI-equipped systems in the classification in the previous section and is conducted on the review items shown in Table 2. For the review, it is necessary to confirm the appropriateness of measures which have to do with compliance with international law, including international humanitarian law, and domestic law, so experts in the legal field and offices in charge will be included as reviewers. As shown in Figure 5, the legal/policy review is important for judging the eligibility of a project (referring to a research and development project), so it will be conducted by reviewing the project implementation office at a meeting where comprehensive decision will be made regarding AI-equipped systems. The review will be conducted before the project commences.

Table 2: Review Items for Requirement A (criteria)

	Confirmation Item
A-1 It is not a project that is non-compliant with international law, including international humanitarian law, and domestic law.	• Is the concept designed to include measures to avoid superfluous injury or unnecessary suffering, considering the balance between military necessity and humanitarian considerations when using the equipment? • Is the concept of the equipment designed to distinguish between military objectives and non-military objectives (civilians and civilian objects) and attack only military objectives (principle of distinction)? • Does the concept include measures to prevent attacks that are expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated (principle of proportionality)?

Table 2: Review Items for Requirement A (Cont.) (criteria)

	Confirmation Item
A-1 It is not a project that is non-compliant with international law, including international humanitarian law, and domestic law.	• Is the concept designed to allow for the implementation of various precautionary measures to prevent indiscriminate attacks and minimize damage to civilians and civilian objects from the selection of military objectives prior to an attack to the execution of the attack (principle of precaution)? • Is the concept of the equipment not such that it cannot be used in accordance with other requirements of international humanitarian law, other applicable international law, and domestic law?
A-2 It is not an autonomous weapon with lethal force that operates completely without human involvement.	• Is the concept designed to allow for an appropriate level of human judgment to intervene and ensure operation within a responsible chain of human command and control? • Is it not the development of an autonomous weapon system with lethal force that operates completely without human involvement?

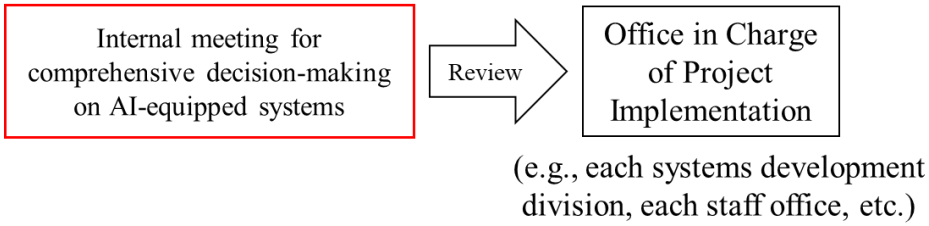

Figure 5: Legal/Policy Review Structure Image

(3) Technical Review

The technical review is conducted to confirm, from a technical perspective, whether prototypes of high-risk AI-equipped systems are equipped with functions that enable them to continue to satisfy Requirement A even at the actual deployment and operation stages, and whether appropriate risk mitigation measures have been implemented. The review is conducted on the review items shown in Table 3. As shown in Figure 6, the technical review will be conducted in a way that the project implementation office will be reviewed at a meeting composed of internal personnel with specialized knowledge of AI technology. Since advanced technical knowledge about ensuring the safety and quality control of AI-equipped systems may be required for the review, opinions from external experts will be sought as necessary. The review will be conducted at milestones in research and development, such as before project commencement, before design approval, before testing commencement, and before operational commencement.

An example of the way of confirming whether the requirements listed in Table 3 are fulfilled in risk management is provided in the Appendix.

Table 3: Review Items for Requirement B (criteria)

	Confirmation Item
B-1 Clarification of Human Responsibility	• Is it designed to allow for appropriate timing and degree of operator involvement and control when using the AI system? • Are the roles of the AI system and the operator clearly defined? • Is the responsibility to be borne by the operator clearly defined?
B-2 Fostering Operator's Appropriate Understanding	• Is it designed to enable operators to become proficient in the behavior, performance range, and operation of the AI system to use it appropriately? • Are measures designed to prevent excessive reliance on the AI system? • Is a mechanism designed to enable the operator to improve the AI system when a malfunction is recognized during monitoring?
B-3 Ensuring Fairness	• Are the fairness requirements for the dataset clarified? • Is it confirmed that biases related to the AI model do not exceed an acceptable level, and is a mechanism designed for improvement when malfunctions are recognized?

Table 3: Review Items for Requirement B (Cont.) (criteria)

	Confirmation Item
B-4 Ensuring Verifiability and Transparency	• Are the processes, methods, data, and algorithms used in the construction of the AI system clarified, and is a mechanism in place capable of later verification of their validity? • Is the person responsible for accountability clearly defined within the research and development organization (including businesses)?
B-5 Ensuring Reliability and Validity	• Are a variety of different metrics used for testing and evaluation to ensure the reliability of the AI system? • Is the use of datasets simulating operational environments considered throughout all the stages from early development to the end of research and development? • Is the design good enough to ensure that AI can be applied without compromising the reliability of equipment connected to AI, assuming its operation? • Is the AI system designed for easy maintenance and management? • Have security control measures for the AI system been considered?
B-6 Ensuring Safety	• Is a safety mechanism designed to reduce the risk of AI system malfunction or serious failure/accident?
B-7 Compliance with International Law and Domestic Law	• Are measures considered to prevent deviations from applicable international laws, domestic laws, and various internal regulations during the operation of the equipment?

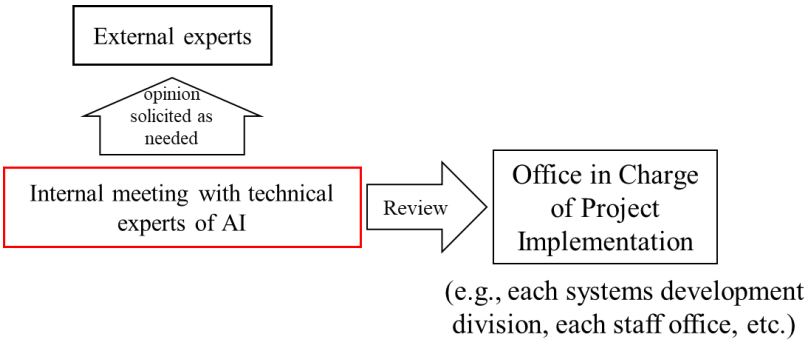

Figure 6: Technical Review Structure Image

Based on Sections 1 to 3, a standard risk management image for high-risk AI-equipped systems is shown in Figure 7.

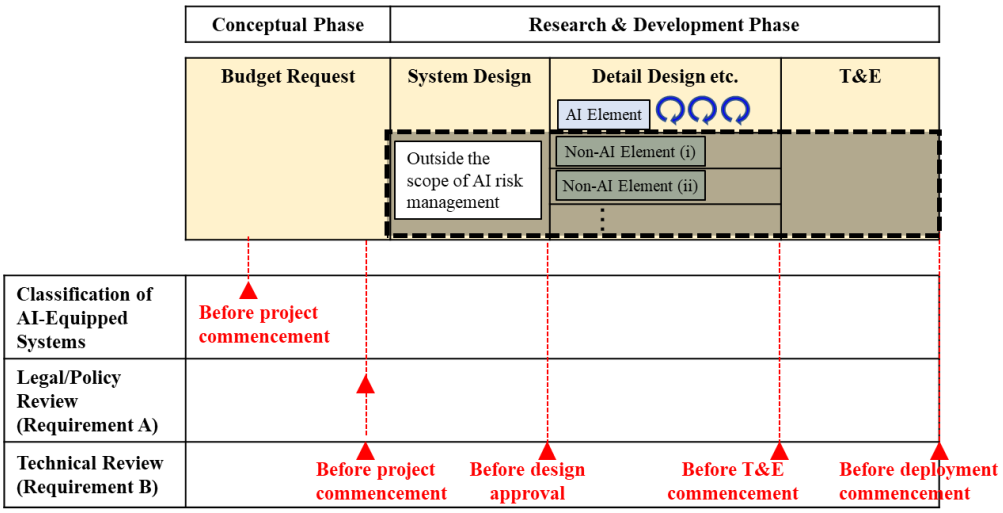

Figure 7: Standard Risk Management Image for High-Risk AI-Equipped Systems

(4) Others

A. Self-Assessment

For items classified as low-risk AI-equipped systems in Section 1, the legal/policy review in Section 2 and the technical review in Section 3 are not required. Risk management will be conducted primarily through self-assessment by the project implementation office. Specifically, this will be done by confirming whether the requirements listed in Table 3 are fulfilled at the appropriate opportunities, such as progress meetings, similar to conventional risk management. Figure 8 shows the image of risk management for low-risk AI-equipped systems.

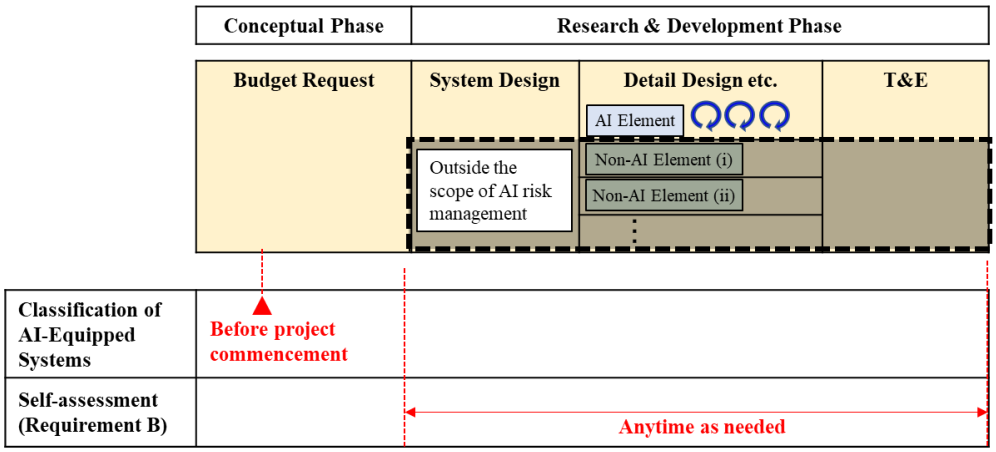

Figure 8: Standard Risk Management Image for Low-Risk AI-Equipped Systems

B. Operator Involvement in Design and Test Evaluation

The performance of AI is significantly influenced by the data used for training. Furthermore, AI performance is also affected by the environment and manner in which it is used. Therefore, it is necessary to test and evaluate AI under a variety of different operational scenarios. Accordingly, it is effective for the Acquisition, Technology & Logistics Agency and each Self-Defense Force to coordinate closely from the conceptual stage and conduct research and development by involving the relevant offices of each staff, units envisioned as future users, units that provide data necessary for AI research and development, and units expected to be responsible for future maintenance and management. It is also effective to establish a framework that allows for the repeated execution of a cycle where prototypes are experimentally deployed to units, operational evaluations are conducted while units experimentally operate the prototypes, and the evaluation results are used to improve or enhance the capabilities of the prototypes. Within this, a phase for experimental operation by units is established, operational evaluations are conducted, and risk management is transferred to the units after sufficient verification of the risks to an acceptable level where maintenance and management by the units is possible. In this way, it is effective to involve operators in the design, test evaluation, and other elements from the earliest possible stage of research and development.

5 Conclusions

The Guideline provides specific guidance for the responsible application of AI in the research and development of equipment. However, as forms of research and development of equipment can be varied, each project design will be flexibly conducted based on the ideas presented in the Guideline as a guide. The situation surrounding AI is changing rapidly, and maintaining rules and procedures once decided may not always be appropriate. Therefore, the Guideline is intended to be a living document and will be updated as appropriate.

For research and development projects that have already commenced at the time of the publication of the Guideline, individual responses will be considered, taking into account the intent of the Guideline and the progress of research and development.

Appendix. An Example of Confirming Whether Requirement B Is Fulfilled Using the RAI Toolkit

1 Purpose

This appendix provides an example of how to confirm whether Requirement B is fulfilled using the RAI Toolkit, intended as reference information for implementers when planning and executing research and development projects related to equipment that applies AI technology.

2 About the RAI Toolkit

The RAI Toolkit, published by the U.S. Department of Defense (Chief Digital and Artificial Intelligence Office (CDAO)), is a tool that can be effectively used for risk management and evaluation in a series of processes from AI development to operation, and most of the tools are industry-standard, open-source options. Furthermore, the RAI Toolkit provides a centralized process that allows AI project implementers to identify, track, and mitigate RAI-related issues (and leverage RAI-related innovation opportunities) throughout the entire responsible AI (RAI) product lifecycle.

3 An Example of Confirming Whether Requirement B Is Fulfilled Using the RAI Toolkit

An example of the RAI Toolkit that can be utilized for confirming whether Requirement B is fulfilled is shown in the following table.

4 References

(1) RAI Toolkit

(2) Responsible AI Toolkit Guide & Front Matter (October 24, 2023)

Table A1: Examples of RAI Toolkit Usable for Requirement B Check Items

No.	RAI Toolkit			Check Item
	RAI Toolkit Tool Name	Tool Description		
1	Trust in Autonomous Systems Test (TOAST)	System Trust Evaluation	A nine-question test to measure how much humans trust a system. (e.g., I understand what the system should do. I understand the system's limitations.) https://www.ida.org/-/media/feature/publications/p/pr/predicting-trust-in-automated-systems---validation-of-the-trust-of-automated-systems-test---toast/d-33088.ashx	B-2 Fostering Operator's Proper Understanding
2	Human-Machine Teaming Systems Engineering Guide	System Design Support	A guide to assist system developers in designing AI that functions in cooperation with human operators. https://www.mitre.org/sites/default/files/2021-11/prs-17-4208-human-machine-teaming-systems-engineering-guide.pdf	
3	FairML	AI Model Fairness Diagnosis/Improvement	A toolkit for analyzing the relationship between AI model prediction results and attributes, such as gender and race, to diagnose fairness, identify factors, and provide improvement methods. Used to improve the fairness of AI model prediction results. https://github.com/adebayoj/fairml	B-3 Ensuring Fairness
4	Tensor Flow Fairness Indicators	AI Model Fairness Evaluation/Improvement	A toolkit for evaluating, improving, and comparing AI model fairness concerns. "Fairness" means that the AI model does not unfairly discriminate based on specific attributes (race, gender, age, etc.). https://github.com/tensorflow/fairness-indicators	

Table A2: Examples of RAI Toolkit Usable for Requirement B Check Items

No.	RAI Toolkit			Check Item
	RAI Toolkit Tool Name	Tool Description		
5	What-If Tool	AI Model Performance and Fairness Evaluation	The What-If Tool (WIT) is a visualization tool that helps understand and analyze machine learning (ML) models, especially those that function as black-box systems in classification and regression tasks. https://github.com/PAIR-code/what-if-tool https://pair-code.github.io/what-if-tool	B-3 Ensuring Fairness
6	Explainer Dashboard	AI Model Interpretability Improvement	This library provides tools to improve the interpretability of AI models. It allows visual analysis of model prediction results and verification of feature importance. https://github.com/oegedijk/explainerdashboard https://explainerdashboard.readthedocs.io/en/latest/	B-4 Ensuring Verifiability and Transparency
7	InterpretML	AI Model Interpretability Improvement	This is an open-source tool that centrally encompasses the latest technologies for machine learning interpretability. This tool provides the ability to create interpretable models and explain black-box systems. It allows an understanding of the overall behavior of a model and the reasons behind individual predictions. https://github.com/interpretml/interpret/ https://interpret.ml/docs/index.html	

Table A3: Examples of RAI Toolkit Usable for Requirement B Check Items

No.	RAI Toolkit			Check Item
	RAI Toolkit Tool Name	Tool Description		
8	Hugging Face Data Card Template	Dataset Transparency Assurance	This shows how to create a dataset card. It helps understand the dataset content, the context of using the dataset, how the dataset was created, and other points users should note. It promotes responsible use and can inform users of potential biases in the dataset. https://huggingface.co/docs/datasets/dataset_card	B-4 Ensuring Verifiability and Transparency
9	Hugging Face Model Card Template	AI Model Transparency Assurance	This is a framework for understanding, sharing, and improving AI models. For each model, it describes its technology and design methods, helping to achieve transparency and auditability. It also helps measure and demonstrate that design procedures are transparent and auditable to demonstrate a proper understanding of the technology. https://huggingface.co/blog/model-cards	
10	Threat Modeling Resource	Threat and Vulnerability Identification/Evaluation and Countermeasure Planning	This is a framework for AI threat modeling. To achieve AI function security, security reviews are conducted through threat modeling, and recommended risk mitigation measures are incorporated based on those reviews. https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml?source=recommendations	B-5 Ensuring Reliability and Validity

Table A4: Examples of RAI Toolkit Usable for Requirement B Check Items

No.	RAI Toolkit			Check Item
	RAI Toolkit Tool Name	Tool Description		
11	EQUI(NE2)	Quantifying Neural Network Uncertainty	This is a tool for visualizing uncertainty in model predictions. It presents a *confidence score* and an *outlier score* for each prediction, allowing evaluation of how much the risk increases when the model operates outside its normal data range. It can evaluate model accuracy and assess model robustness to data bias and clarify the applicable range of AI. https://github.com/mit-ll-responsible-ai/equine	B-5 Ensuring Reliability and Validity
12	IBM Adversarial Robustness 360Attacks	AI Security Enhancement	This Python library for AI security can generate adversarial attacks, train robustness to enhance AI reliability (AI security), and calculate attack success rates to measure and demonstrate AI reliability (AI security). It provides tools for users to protect and evaluate machine learning models and applications from the adversarial threats of evasion, poisoning, extraction, and inference. It supports the AI frameworks of TensorFlow, Keras, PyTorch, scikit-learn, and XGBoost and handles all data types, such as images, tables, audio, and video, and the AI tasks of classification, object detection, speech recognition, generation, and authentication. https://github.com/Trusted-AI/adversarial-robustness-toolbox/tree/main/art/attacks https://github.com/Trusted-AI/adversarial-robustness-toolbox/wiki/ART-Attacks	

Table A5: Examples of RAI Toolkit Usable for Requirement B Check Items

No.	RAI Toolkit			Check Item
	RAI Toolkit Tool Name	Tool Description		
13	Drift Tools	Supporting AI Function Reliability and Governance	This Python library helps implement AI reliability (effectiveness) and the ability to detect unintended outcomes by incorporating Drift Tools algorithms into AI functions to detect out-of-distribution inputs where AI function performance is not guaranteed. It specializes in outlier detection, adversarial detection (the process of detecting and defending against attacks on machine learning models), and drift detection (the process of detecting changes that occur between the statistical distribution of machine learning model training data and the distribution of data actually encountered). https://github.com/SeldonIO/alibi-detect https://docs.seldon.io/projects/alibi-detect/en/stable/	B-2 Fostering Operator's Proper Understanding B-5 Ensuring Reliability and Validity
14	Python Outlier Detection (PyOD)	Detecting Anomalies in Multivariate Data	PyOD (Python Outlier Detection) is a comprehensive and scalable Python library for detecting anomalies in multivariate data. This library provides a wide range of algorithms to meet a variety of needs from small-scale projects to large datasets. https://github.com/yzhao062/pyod https://pyod.readthedocs.io/en/latest/	B-6 Ensuring Safety

【Explanation】

B-2's question: "Is a mechanism designed to enable the operator to improve the AI system when a malfunction is recognized during monitoring?" addresses the risk of new malfunctions being discovered by operators as the AI system is used in a real environment. The Drift Detection function in the RAI Toolkit's Alibi Detect is an open-source tool that helps monitor and maintain AI performance. Other references that can be consulted include the MITRE Human-Machine Teaming Systems Engineering Guide, which helps design AI that functions in cooperation with operators, the Institute for Defense Analyses (IDA) Trust in Autonomous Systems Test (TOAST), which contains nine questions for determining how well humans trust a system, and the U.S. Department of Defense CDAO's Human Systems Integration (HSI) T&E[*1] literature.

For ensuring fairness in B-3, the RAI Toolkit's What-If Tool (WIT) is an open-source visualization tool that helps investigate model performance and fairness across different subsets of data. Additionally, the RAI Toolkit's FairML and TensorFlow Fairness Indicators are useful for diagnosing and improving fairness.

For ensuring transparency in B-4, examples include using Data Cards, which are templates helpful for understanding dataset content, creation methods, and other points users should note, and Model Cards, which are templates helpful for understanding, sharing, and improving AI models. Furthermore, the RAI Toolkit's InterpretML is an open-source tool that centrally encompasses the latest technologies for machine learning interpretability. It helps understand the overall behavior of a model and the reasons behind individual predictions. Similarly, the RAI Toolkit's Explainer Dashboard library also helps improve interpretability by analyzing model prediction results and checking feature importance.

*1:https://cdao.pages.jatic.net/public/guidance/human-systems-integration/

For the various metrics in B-5, in addition to the most intuitive metric, Correctness (accuracy etc.), the U.S. Department of Defense's CDAO presents such perspectives as Bias (selection or bias of training data), Resilience (function recovery), Uncertainty (uncertainty against accuracy, confidence, etc.), Representativeness (representativeness of training data assuming an operational environment, etc.), Latency (processing speed), Explainability (explainability), Robustness (resistance to adversarial attacks and noise, etc.), and Drift (data or model drift) in its AI Model T&E[*2] literature. The RAI Toolkit's EQUI(NE2) is an open-source tool that can quantify not only the accuracy of neural network output but also its uncertainty. B-5's question: " Is the use of datasets simulating operational environments considered throughout all the stages from early development to the end of research and development?" and "Is the design good enough to ensure that AI can be applied without compromising the reliability of equipment connected to AI, assuming its operation?" are based on the U.S. Department of Defense's CDAO's Operational T&E (OT&E)[*3] and Systems Integration (SI) T&E[*4] literature. The RAI Toolkit's Adversarial Robustness Toolbox (ART) is an open-source security tool that helps protect and evaluate against the adversarial threats of evasion, poisoning, extraction, and inference and is useful for implementing security control measures.

In B-6, for ensuring safety, the Python library PyOD (Python Outlier Detection), which provides a wide range of algorithms for detecting outliers, is useful for designing safety mechanisms.

By using the Requirement B check items as a framework for discussion and proceeding with research and development after considering the appropriate risk countermeasures based on the answers, it becomes easier to confirm whether there are any omissions in the necessary processes and whether there are any potential problems in the construction of the AI system. At this time, effective risk management can be achieved by soliciting opinions on risk assessment from experts who are familiar with past failures in research and development of AI systems and new technology information and who sufficiently understand the characteristics and vulnerabilities of AI. Referring to such databases as the RAI Toolkit's AI Incidents Database can also provide insights into risk identification and countermeasures from past experiences.

*2: https://cdao.pages.jatic.net/public/guidance/model/

*3: https://cdao.pages.jatic.net/public/guidance/operational/

*4: https://cdao.pages.jatic.net/public/guidance/systems-integration/