OECD AI Transparency Report

Organization: Preferred Networks (JP) **Reporting Period:** Q2 2025 **Published:** April 11, 2025

Section 1 – Risk identification and evaluation

a. How does your organization define and/or classify different types of risks related to AI, such as unreasonable risks?

We define and classify risks based on established guidelines such as the AI Guidelines for Business and QA4AI.

b. What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle?

We have developed our AI Quality Guidelines for AI system development and implementation based on our internal AI governance guidelines. By following these guidelines, we ensure to understand the implementation status of each project.

c. Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model's/system's fitness for moving beyond the development stage?

We are considering implementing red team testing for our LLM products.

d. Does your organization use incident reports, including reports shared by other organizations, to help identify risks?

Yes

e. Are quantitative and/or qualitative risk evaluation metrics used and if yes, with what caveats? Does your organization make vulnerability and incident reporting mechanisms accessible to a diverse set of stakeholders? Does your organization have incentive programs for the responsible disclosure of risks, incidents and vulnerabilities?

Yes, we use the metrics provided in the AI Guidelines for Business and carefully implement them throughout our product development process.
We also provide a contact point for users, enabling us to gather diverse opinions.
We do not have such an incentive program.

f. Is external independent expertise leveraged for the identification, assessment, and evaluation of risks and if yes, how? Does your organization have mechanisms to receive reports of risks, incidents or vulnerabilities by third parties?

Red teaming testing will be conducted by an external organization.
There is a dedicated contact point on our website for reporting security and privacy issues.

g. Does your organization contribute to the development of and/or use international technical standards or best practices for the identification, assessment, and evaluation of risks?

We use multiple references such as the AI Guidelines for Business, QA4AI, and OWASP LLM TOP10.

h. How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks?

We have established an AI Governance Promotion team as part of our company-wide risk management initiative led by our CEO.

Any further comments and for implementation documentation

No answer provided

Section 2 – Risk management and information security

a. What steps does your organization take to address risks and vulnerabilities across the AI lifecycle?

Company guidelines have been established based on The AI Guideline for Business, QA4AI, and others.
A company-wide risk management system has been implemented.

b. How do testing measures inform actions to address identified risks?

For general risks, we direct the necessary countermeasures as part of company-wide risk management.
In addressing risks related to AI systems, it is our plan to evaluate them with AC (Answer Carefully Dataset), BBQ (a hand-built bias benchmark for question answering), OWASP LLM TOP10 among others, and provide feedback for AI systems as part of alignment.

c. When does testing take place in secure environments, if at all, and if it does, how?

Tests are carried out in a test environment that does not affect the production environment.

d. How does your organization promote data quality and mitigate risks of harmful bias, including in training and data collection processes?

Internal guidelines have been formulated based on The AI Guideline for Business and QA4AI, and guidelines have been established for products accordingly.
The introduction of tests to reduce bias is being considered (AC :Answer Carefully Dataset, BBQ: a hand-built bias benchmark for question answering)).

e. How does your organization protect intellectual property, including copyright-protected content?

The PFN AI Quality Guidelines clearly state the appropriate handling of data, including training data.
In addition, red teaming tests are conducted as necessary.

f. How does your organization protect privacy? How does your organization guard against systems divulging confidential or sensitive data?

The PFN AI Quality Guidelines clearly state the appropriate handling of data, including training data.
Best practices for system security are provided, and security checks are carried out to confirm that the system is ready for release.

**g. How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?
i. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems?
ii. How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and work with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?
iii. What is your organization's vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?
iv. How often are security measures reviewed? v. Does your organization have an insider threat detection program?
**

Firstly, physical risks are identified, and the requirements for reducing these are presented to the AI side.

i. We have established policies and guidelines and defined the procedures for implementing these in our products.
Security testing is mandatory for products.
ii. Security measures are implemented as necessary after evaluation during security testing for each product, as they vary depending on the product implementation.
iii. A contact point for reporting security and privacy issues has been set up on the website.
Vulnerability information provided by CERT organizations and others is collected, and necessary measures are taken.
iv. The policy level is scheduled to be reviewed annually, and the implementation level is scheduled to be reviewed sequentially.
v. Internal threats are dealt with by monitoring the status of system log acquisition, authority settings and changes to authority, and other significant indicators.

h. How does your organization address vulnerabilities, incidents, emerging risks?

Before release: A company-wide risk assessment is conducted for general risks.
For products, red teaming tests and third-party evaluations are conducted as necessary.
After release: The AI Governance and Security Team continuously monitors vulnerabilities and incidents.
A contact point is set up on the website to receive information from third parties.

Any further comments and for implementation documentation

No answer provided

Section 3 – Transparency reporting on advanced AI systems

**a. Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?
i. How often are such reports usually updated?
ii. How are new significant releases reflected in such reports?
iii. Which of the following information is included in your organization's publicly available documentation: details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights; assessments of the model's or system's effects and risks to safety and society (such as those related to harmful bias, discrimination, threats to protection of privacy or personal data, fairness); results of red-teaming or other testing conducted to evaluate the model's/system's fitness for moving beyond the development stage; capacities of a model/system and significant limitations in performance with implications for appropriate use domains; other technical documentation and instructions for use if relevant.
**

Technical documents are provided for specific customer systems as part of the delivery.
The terms of use and privacy policy are clearly stated for systems that do not specify customers.
i. Documents are updated as needed, typically when changes in specifications require revisions.
ii. We clearly state their importance along with any necessary responses.
iii. The Terms of Use, Privacy Policy, and/or other documents state any information that conflicts with users' rights.
Although the details of risk assessment are not always clearly stated, the above documents state the points that users should consider.
Test results are not disclosed unless required by contract.
Significant performance limitations are stated in the Terms of Use, User Guide, and/or other documents.
Other information is disclosed and published as necessary.

b. How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

We publish it as blog posts, white papers, and academic papers:
https://tech.preferred.jp/en/

c. Does your organization disclose privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

Our privacy policy is published on our website: https://www.preferred.jp/en/policy/

d. Does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

Yes. For example, we have published a blog post and arxiv papers containing such information for our PLaMo 100B model:
https://tech.preferred.jp/ja/blog/plamo-100b/
https://arxiv.org/abs/2410.07563v2
https://huggingface.co/pfnet/plamo-100b

e. Does your organization demonstrate transparency related to advanced AI systems through any other methods?

We demonstrate transparency by participating in various committees and initiatives related to AI safety and sharing our company's initiatives.

Any further comments and for implementation documentation

No answer provided

Section 4 – Organizational governance, incident management and transparency

a. How has AI risk management been embedded in your organization governance framework? When and under what circumstances are policies updated?

AI Governance Promotion has been established as part of company-wide risk management directly under the Board of Directors, and is being implemented in cooperation with the company-wide risk management department.
The policy is to be reviewed annually (scheduled).

b. Are relevant staff trained on your organization's governance policies and risk management practices? If so, how?

We stipulate in our guidelines that, at a minimum, annual training must be carried out.

c. Does your organization communicate its risk management policies and practices with users and/or the public? If so, how?

The information is published as our AI policy on our website.
https://www.preferred.jp/en/company/aipolicy/

d. Are steps taken to address reported incidents documented and maintained internally? If so, how?

The response to IT-related incidents is stipulated in our documentation, and we follow these procedures when responding to AI-related incidents as well.

e. How does your organization share relevant information about vulnerabilities, incidents, emerging risks, and misuse with others?

If an incident occurs in our system, it will be reported to the customer following IT incident response procedures, and if necessary, it will be shared through blogs, papers, and/or other media.

f. Does your organization share information, as appropriate, with relevant other stakeholders regarding advanced AI system incidents? If so, how? Does your organization share and report incident-related information publicly?

At present, we have not experienced any incidents involving advanced AI systems.

In the event of an incident, services based on contracts will share the incident in accordance with the contract, and in the case of services with terms and conditions, communication will be carried out via release notes or other reasonable means.

g. How does your organization share research and best practices on addressing or managing risk?

We share our best practices by participating in government, university, and other committees and research projects.

h. Does your organization use international technical standards or best practices for AI risk management and governance policies?

We believe our organization follows international best practices, as our risk management and governance policies are based on the AI Guidelines for Business.

Any further comments and for implementation documentation

No answer provided

Section 5 – Content authentication & provenance mechanisms

a. What mechanisms, if any, does your organization put in place to allow users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

Our AI Quality Guidelines clearly define that AI products should be recognizable as AI. Additionally, we ensure this recognizability is considered when formulating terms of use and other related documents, through legal checks.

b. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems? If yes, how? Does your organization use international technical standards or best practices when developing or implementing content provenance?

Our system clearly states through the service interface that content is generated by AI.

Any further comments and for implementation documentation

Our system clearly states through the service interface that content is generated by AI.

Section 6 – Research & investment to advance AI safety & mitigate societal risks

a. How does your organization advance research and investment related to the following: security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems?

We have formulated AI quality guidelines for product development based on the OECD AI Guidelines and the METI AI Guidelines for Business. These guidelines are followed as procedures for implementing these standards in our products.
We are also considering evaluations using AC (Answer Carefully Dataset), BBQ (a hand-built bias benchmark for question answering).

b. How does your organization collaborate on and invest in research to advance the state of content authentication and provenance?

At present, we are in the research stage and have not yet undertaken specific initiatives for our products.

c. Does your organization participate in projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools?

We serve as members of several government committees, such as METI's Study Group of AI Guidelines for Business, the Digital Agency's Data Security Working Group, the Cabinet Office's Study Group for Intellectual Property in the Age of AI, and MIC's G7 Reporting Framework. We also participate in Japan Network Security Association (JNSA), a security industry organization.

d. What research or investment is your organization pursuing to minimize socio-economic and/or environmental risks from AI?

We are developing low-power AI chips and have been ranked #1 in the GREEN 500 several times.

Any further comments and for implementation documentation

No answer provided

Section 7 – Advancing human and global interests

a. What research or investment is your organization pursuing to maximize socio-economic and environmental benefits from AI? Please provide examples.

We developed AI from the chip level and implemented it as a cluster system to reduce the power consumption issues associated with the use of AI. Our supercomputers have achieved multiple top rankings in the GREEN 500.
To improve the performance of LLMs in the context of the Japanese language and culture, we have developed PLaMo as a foundation model and launched it as a product.
Furthermore, our company is actively engaged in various socio-economic and environmental initiatives. These include participating in research related to climate prediction, developing human body models as digital twins, developing and commercializing autonoous driving systems, and providing educational support.
Also as part of our efforts to promote diverse working styles among our employees, we have adopted remote work.

b. Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities, limitations and impacts of advanced AI systems? Please provide examples.

Providing digital educational content such as Playgram, training courses related to AI, and AI educational content for elementary school students.

c. Does your organization prioritize AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals? Please provide examples.

We believe that making the real world computable through AI will contribute to society at large. To keep environmental costs down, we are working on sustainable development in all areas of our business by developing from the chip level.

d. Does your organization collaborate with civil society and community groups to identify and develop AI solutions in support of the UN Sustainable Development Goals and to address the world's greatest challenges? Please provide examples.

Through participation in government and other committees, we cooperate with citizen groups and consumer groups.

Any further comments and for implementation documentation

No answer provided