

OECD AI Transparency Report

Organization: Karavela (FRA) Reporting Period: Q3 2026 Published: October 2, 2026

Section 0 – Your role in the AI system lifecycle

Responding as (please select all that apply):

- Model developer (/provider)
- Other (please specify): Data collection

If any of the fields above are selected, please indicate your organization size:

Small enterprise: 10-49 employees

Section 1 – Risk Identification and Evaluation

1. How does your organization define and/or classify different types of risks related to AI?¹

- Maintain an internal taxonomy, classification system, and/or set of risk area definitions

How does your organization define thresholds at which severe risks, unless adequately mitigated, may be considered "unreasonable"? Please describe the framework(s) most relevant to your organization's role and context. If your organisation does not use formal thresholds, you may describe instead the policies, decision criteria, or governance processes you apply when assessing severe AI risks. You may also describe: • Specific examples of contexts in which your models or systems could, in your assessment, pose unreasonable or intolerable risk; • How mitigations are graduated by capability or propensity level; • How your organisation considers the specificities of different AI technologies, such as generative AI and agentic AI, including any specific risk assessment or governance mechanisms applied to agentic AI capabilities.

Karavela develops AI models using functional MRI data to study brain function and support research into neurological and psychiatric conditions. In this context, the most relevant severe risks concern participant and patient safety, the confidentiality of brain and health data, mental privacy, and inappropriate reliance on model outputs.

Karavela applies the risk methodology of its ISO/IEC 27001 information security management system. Each risk is scored as probability × impact on a scale of 1 to 16. Scores of 1–4 are acceptable, scores of 5–11 are acceptable only with a treatment plan, and scores of 12–16 are unacceptable and require immediate action. AI-specific risk scenarios are included in our risk register and assessed on the same scale. This methodology is complemented by human-subject research ethics, data-protection requirements, and scientific and clinical validation. We assess the potential severity, likelihood, scale

and reversibility of harm; the vulnerability of affected individuals; the intended use and foreseeable misuse; and the effectiveness of available safeguards.

We would consider a use unreasonable where a credible pathway to severe harm remains insufficiently controlled. Examples include:

- Using an insufficiently validated model to make consequential diagnostic or treatment decisions.
- Inferring or disclosing sensitive cognitive or health information outside the authorised research purpose
- Enabling coercive or discriminatory uses of brain-derived information, including employment or insurance decisions.
- Presenting probabilistic neural decoding outputs as reliable evidence of an individual's thoughts, intentions or beliefs.

Safeguards should increase with demonstrated capabilities, access to sensitive data, deployment scale and decision-making authority. Uses with potential clinical consequences require stronger validation across participants, sessions and acquisition settings as well as assessment of failure modes and subgroup performance. Where residual risk cannot be adequately mitigated, the relevant use should be restricted.

For generative neural decoding, a particular concern is that plausible generated content may not faithfully reflect the recorded brain activity. Fluency must therefore not be treated as evidence of decoding accuracy. Evaluation and communication should distinguish measured predictive performance from interpretation.

2. Both before and after deployment, what practices does your organization use to identify and evaluate vulnerabilities, incidents, emerging risks and misuse throughout the AI system lifecycle?

- Publicly available benchmarks and measurement tools
- Internal benchmarks, standard metrics (e.g., frequency, severity scores)
- Structured techniques to explore the behavior or performance of a model and/or application, such as model elicitation or bottleneck assessments
- Qualitative assessments (e.g., expert judgment, narrative scoring)
- Internal model, system, or application-independent research initiatives
- External model, system, or application-independent research initiatives

Please briefly describe your organization's decision-making process for deploying an AI model or system including: who has decision authority; what evidence and criteria are required (e.g., performance, safety/security, robustness, fairness, privacy, compliance, independent external testing, evaluation, validation and verification (TEVV) or assurance); contextual caveats or limitations required in interpreting risk evaluation metric results; what approvals/sign-offs are needed; and what safeguards (e.g., mitigations, monitoring, or phased release) must be in place to move into deployment.

Karavela has not yet deployed an AI model or system. Our models are currently developed and evaluated in a research setting, and a formal deployment approval process has not yet been

established.

Before any deployment, we expect decision authority to rest with company leadership, informed by scientific and technical leads and, where relevant, clinical, legal and data-protection expertise. Required evidence would be proportionate to the intended use and include performance, robustness across participants, sessions and MRI acquisition settings, relevant subgroup assessments, privacy and security evaluations, and compliance with applicable requirements. Any external validation or independent testing would be reported explicitly.

Results would be interpreted within the limits of the evaluated populations, datasets and tasks.

Research performance alone would not establish clinical suitability, and neural decoding outputs would not be treated as definitive evidence of an individual's thoughts or intentions.

Deployment would require documented internal approval and any applicable external authorisations.

Expected safeguards include a controlled initial release, restricted access, clear intended-use limitations, appropriate human oversight, incident monitoring and the ability to suspend use. These arrangements remain prospective and would be formalised before deployment.

Following deployment, how do you use incident reports — including those shared by other organizations — to identify emerging issues, update risk assessments, and trigger corrective actions.

Karavela has not yet deployed any AI models or systems, so post-deployment incident management is not yet applicable. Before deployment, we plan to establish a process to review internal and external incident reports, update risk assessments, and trigger appropriate corrective actions.

3. Does your organization have mechanisms to receive and/or share reports of risks, incidents or vulnerabilities by/with third parties?

Yes, both receiving and sharing

If yes, please indicate which stakeholders reporting mechanisms may be accessible to, recognising specifics will depend on the nature of the report and risk level. For each, indicate whether access is on a standing basis (Always), activated based on context (Ad hoc/as relevant), or not currently accessible.

	Always	Ad hoc / as relevant	Not currently accessible
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities (e.g., public reporting offices for security incidents or data breaches)		☐	
Supervisory authorities (e.g., data protection authorities)		☐	
Industry partners or industry fora		☐	
Users (end users, enterprise customers, developers using APIs)		☐	

	Always	Ad hoc / as relevant	Not currently accessible
Business customers / deploying organizations (if different from “users”)		☐	
Supply-chain partners (cloud, data providers, etc.)		☐	
Academia / researchers			☐
Independent auditors		☐	
Standards bodies / multi-stakeholder initiatives (OECD.AI, ISO/IEC, IEEE, etc.)			☐
Specific civil society organizations			☐
Media / general public (people or communities indirectly impacted)			☐
Other			☐

Please describe how your organisation determines when and how reporting mechanisms are made accessible to different stakeholders (e.g. based on legal or regulatory requirements, internal policies, or risk level). Please also provide public URLs or details on how to submit reports. If none, write “Not applicable”

Karavela receives reports of security incidents and vulnerabilities from employees, partners and other third parties through a dedicated address, security@karavela.ai. Staff must report any incident within 24 hours. Each report is assessed for severity and handled under our Security Incident Response Policy. Where the law or a contract requires it, we inform the affected partners, individuals and authorities, including the CNIL for personal data breaches.

How does your organization share relevant information as appropriate?

- Reporting to regulatory authorities as relevant

Please describe what type of information is shared.

Where the law or a contract requires it, we inform the affected partners, individuals and authorities, including the CNIL for personal data breaches.

4. If your organization has programs to incentivize risk(s) disclosure, incidents and systemic vulnerabilities, please share what type of programs you use?

- Not applicable

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

Uses

If contributes or uses, please select all that apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001, ISO/IEC 23894, ISO/IEC 42005)
- Industry-led best practices or codes of conduct

6. Does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks that could have broad and potentially cascading socio-economic impacts? Please indicate which stakeholders you collaborate with, and whether collaboration is on a standing basis (Always-on), activated based on context (Ad hoc/as relevant), or there is no current collaboration.

	Always-on	Ad hoc / as relevant	No current collaboration
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities / regulators / AI safety institutes / other public sector bodies and inter-governmental coordination structures		☐	
Industry partners or industry fora		☐	
Users (end users, enterprise customers, developers using APIs)		☐	
Business customers / deploying organizations (if different from “users”)		☐	
Supply-chain partners (cloud, data providers, etc.)		☐	
Academia / researchers	☐		
Independent auditors		☐	
Standards bodies / multi-stakeholder initiatives (OECD.AI; ISO/IEC; IEEE; AIID etc.)		☐	
Independent civil society organizations (e.g., for evaluations, audits, or policy review)			☐
Media / general public (people or communities indirectly impacted)		☐	
Other(s) (please add any other relevant stakeholder groups):			☐
Not applicable			

Please describe examples of how your organization collaborates. Examples may include: participating in multi-stakeholder or industry initiatives; partnering with academia, civil society, or the public sector; establishing internal or joint governance structures with external experts or in partnership with other efforts to share knowledge and jointly address risks.

Karavela collaborates with academic researchers, hospitals and clinical investigators on functional MRI data collection, study design and scientific validation. Our multicentre research involves partners including Fondation Rothschild, Paris Brain institute and INT, with investigator input and independent ethics review supporting participant protection and responsible data use. We also plan to share research datasets and findings with the scientific community to support reproducibility, external evaluation and a better understanding of neural decoding capabilities and limitations.

Any further comments and implementation documentation.

No answer provided

Section 2 – Risk Management and Information Security

7. Across the AI system lifecycle, what steps does your organization take to manage risks and vulnerabilities? Please indicate which specific practices, tools or metrics your organization uses at each lifecycle stage. You may refer to tools or metrics present in the OECD.AI Catalogue of Tools and Metrics or suggest additions. Respondents may complete only the lifecycle stages relevant to their organisation's role. Stages that do not apply may be left blank.

Plan and design

Collect and process data

Build, integrate and interpret model

Verify and validate

Deploy

Operate and monitor

Please provide a consolidated description of your organisation's approach to risk management across the AI system lifecycle, including how the practices identified above operate together and any cross-cutting governance, review or escalation processes. Please also indicate how often these practices are reviewed and what types of events trigger reviews and updates. Where relevant, please distinguish between different types of AI systems (e.g. generative or agentic AI).

Karavela's AI systems remain at the research stage, with no deployment yet. Our approach combines research ethics oversight, data protection and scientific evaluation, with significant concerns escalated to leadership and relevant experts. Risks are reassessed at least once a year, and whenever capabilities, datasets or intended uses change or new evidence of risk emerges. Risk treatment status is reported to management twice a year, and policies are reviewed annually. Generative decoding requires particular attention to misleading outputs and mental privacy. Any future deployment or agentic capability would require additional assessment and safeguards.

Please describe any measures used to ensure data quality and fairness across the data collection and curation stage (e.g., dataset audits, filtering training data to remove harmful content or personal

information, fine-tuning on curated data, use of moderation tools, assessment of limitations to model generalisability).

Karavela uses standardised acquisition protocols and quality-control workflows to assess functional MRI data quality and identify acquisition artefacts. Research data are pseudonymised, with access restricted to authorised personnel. Curation focuses on signal quality and suitability for the research purpose. We consider participant characteristics, scanner and site differences, and sampling limitations when interpreting results; performance on research datasets does not establish generalisability across populations or clinical settings.

8. What security measures are used in testing AI systems?

- Isolation of staging environments (from production / operational environments)
- Restrictions on outbound Internet or API access, including a list of permitted domains or APIs
- Human-in-the-loop controls for higher-risk test cases
- Audit logging and monitoring

8.A. Please indicate, if relevant, the team or role responsible for overseeing secure testing and whether this responsibility is shared across multiple teams and/or functions.

The CISO oversees secure testing, supported by the security team.

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

9.A. Safeguards that promote lawful data use and transparency in AI training:

- Other(s) (please specify): Karavela is mostly training on his own proprietary dataset, acquired specifically for model training
- Maintain a copyright policy that addresses data sourcing, content provenance, safeguards, and organizational roles and responsibilities
- Refrain from the unlawful circumvention of technological protection measures in data sourcing
- Consult and/or retain an intellectual property attorney to protect, enforce, and comply with, intellectual property rights.

9.B. Safeguards that reduce the risk of dissemination of outputs that may infringe copyright:

- Not applicable

10. How do you protect privacy and prevent your systems from disclosing confidential or sensitive data across the AI lifecycle?

- Use security controls, including encryption at rest and in transit
- Apply Privacy-Enhancing Technologies (PETs) where applicable (e.g. differential privacy, federated learning, synthetic data generation, homomorphic encryption)

Describe key privacy measures and data-protection standards applied.

Karavela's approach is guided by the GDPR and applicable French requirements for health data and human-subject research. Key measures include informed participant consent, purpose limitation, data

minimisation, pseudonymisation, restricted access, and secure storage and transfer. Data use and sharing must remain consistent with research authorisations, participant information and applicable legal requirements.

A Data Protection Officer oversees compliance and maintains the register of processing activities.

Health data is hosted in France by a provider certified for health data hosting (HDS).

11. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems across the AI system lifecycle? Respondents may complete the checkbox subsections in 11.A–11.D or describe their overall approach under 11.E.

11.A. Governance, policy and assessments

- Maintain or update security policies aligned with recognized standards (e.g., NIST CSF, ISO/IEC 27001, SOC 2, CIS Critical Security Controls), and review them regularly in response to emerging risks or regulatory guidance
- Implement incident response, disaster recovery, or business continuity plans, including escalation procedures for high-severity incidents
- Train staff on cybersecurity best practices and AI-specific threats, including how to recognize and respond to misuse or unintended outcomes of advanced AI systems

11.B. Technical and operational controls

- Apply multi-factor authentication (MFA) and role-based access controls for advanced AI systems
- Encrypt AI assets (datasets, models, configurations) in storage and in transit as appropriate

11.C. Protection of sensitive or high value AI assets

- Limit access to proprietary AI assets based on role or need-to-know
- Store model weights, datasets, and algorithms in secure, access-controlled environments

11.D. Insider risk protection detections

- Provide staff training on cybersecurity best practices and AI-specific threats
- Monitor for unauthorized internal access or suspicious activity
- Implement insider-risk escalation or response procedures
- Conduct pre-employment screening or role-sensitive vetting

11.E. Please describe the main security measures your organization applies and indicate which are specific to particularly sensitive AI assets, where relevant.

Karavela's security measures are defined in an information security management system aligned with ISO/IEC 27001. They include single sign-on with multi-factor authentication, role-based access restricted to need-to-know, encryption, pseudonymisation of research data, and audit logging. The most sensitive assets are brain data, which have the strictest access controls. Model weights, checkpoints and training artefacts are classified as confidential and stored only in access-controlled environments. Our experiment-tracking platform requires each user to authenticate through our SSO using an individual account.

12. 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?

- Reporting to regulators / authorities when relevant
- Internal vulnerability disclosure receipt and tracking processes
- Investigation of reported or identified risks and vulnerabilities
- Development of patches or other mitigations for confirmed risks or vulnerabilities
- Testing of patches or other mitigations
- Scanning for and mitigation of similar risks or vulnerabilities
- Defined escalation criteria and follow-on actions for identified vulnerabilities (e.g., for external reporting or information sharing)

13. What post-deployment practices does your organization use to monitor and respond to vulnerabilities, incidents, emerging risks and misuse of advanced AI systems?

- Other(s) (please specify): No deployment yet

14. Do you develop or deploy agentic AI systems (systems that can autonomously plan/act or use tools)?⁴

No

Please provide any further comments and implementation documentation.

No answer provided

Section 3 – Transparency Reporting on Advanced AI Systems

15. Does your organization publish public reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems? Are these clear and understandable by non-technical audiences?

No

16. 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?

- Public documentation

Describe how evaluation results are shared and with which stakeholders (e.g. other organizations, governments, civil society and academia, etc). You may also indicate whether non-technical or plain-

language summaries are provided for broader audiences.

Karavela plans to disseminate findings broadly through scientific publications and dataset releases, documenting methods, performance and limitations while protecting participant confidentiality. Public communication aims to make findings accessible without overstating model capabilities or clinical readiness.

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

No

18. How 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?

- Public documentation
- Protected information sharing

Please provide links to any publicly available information.

<https://karavela.ai/en/studies/cephalonauts-one>

<https://karavela.ai/en/studies/cephalonauts-one/transparency>

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

No

20. Does your organization publish statistical insights derived from anonymised, aggregated usage data to advance understanding of AI adoption and its economic impacts in areas including occupations and tasks by region, sector and country? To this end, do you contribute to international cooperation efforts?

No

Please estimate the total staff time and annual expenses associated with conducting compliance activities and include any further comments and implementation documentation (optional).

No answer provided

Section 4 – Organizational Governance, Incident Management, and Transparency

21. How has AI risk management been embedded in your organization's governance framework?

- Executive oversight

Please describe how AI risk management is integrated into governance structures, including: which roles or bodies have advisory versus decision-making authority; which roles or bodies have final accountability for AI risk-related decisions; and how often policies are reviewed or updated (e.g., following major incidents, regulatory changes, or material updates to systems).

AI risk management is part of our ISO/IEC 27001 management system. The CEO holds final accountability for accepting and treating risks, informed by scientific and technical leads and by clinical, legal and data-protection experts. A Management Review Committee (CEO, CTO and CISO) meets at least once a year and approves the risk assessment, the risk treatment plan and the residual risks. A Security Steering Committee (CEO, CTO, CISO and DPO) meets monthly in an advisory and operational role, following up on risks, incidents and actions. Independent ethics committees provide external oversight of human-subject research. Policies are reviewed every year, and also after material changes to models, data or intended uses, significant incidents, new risk evidence or regulatory changes. Risk treatment status is reported to management twice a year.

22. Are relevant staff trained on your organization's governance policies and risk management practices?

Yes

If yes, describe training methods (e.g., onboarding modules, regular sessions, and confidential channels to report concerns).

Security awareness training is provided to all staff when they join and on an annual basis, with participation recorded. New staff must read and formally acknowledge our security policies, including the Acceptable Use Policy and its rules on the use of AI tools. Guidance specific to our research covers research protocols, data confidentiality, secure handling of sensitive data and model limitations. Staff report security concerns and incidents through our dedicated security-incident reporting channel.

23. Are steps taken to address reported incidents documented and maintained internally?

Yes

If yes, describe how incident records are kept and used for follow-up (e.g., incident logs, root-cause analysis reports, internal tracking systems). You may indicate whether and how you define escalation thresholds for when incidents must be surfaced to senior leadership or external parties.

All security incidents and the actions taken to resolve them are recorded in our security incident registry and handled in accordance with our Security Incident Response Policy. Each incident is classified by severity. Major incidents require a root cause analysis and a post-incident review, and their lessons feed into our risk assessment. Incident records are kept for five years.

24. Does your organization share research and best practices on addressing or managing risk?

No

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

Yes

If yes, please indicate which standards or frameworks apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001)

If yes, describe which international technical standards or best practices your organization uses.

Karavela runs an information security management system aligned with ISO/IEC 27001. Our AI governance documents, including our AI Security Policy, belong to this system: each has an owner, a version history and an annual review. AI-related risks are assessed with the same risk methodology as other information security risks, and AI-specific risk scenarios are included in our risk register.

Any further comments and implementation documentation.

No answer provided

Section 5 – Content Authentication & Provenance Mechanisms

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

- Other (specify): No tool deployed yet

Provide any additional details or context on how or when users are informed they are interacting with advanced AI systems (e.g. if such mechanisms are planned, in testing, or applied only to certain products).

No answer provided

27. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems?

Not applicable

Any further comments and implementation documentation.

No answer provided

Section 6 – Research & Investment to Advance AI Safety & Mitigate Societal Risks

28. Does your organization advance research and investment in the following areas?

- Data quality and representativeness
- System performance and outcomes (e.g. accuracy, consistency)

If yes, please describe key research or investment initiatives in the areas selected above (e.g., dedicated safety teams, funding of research programs, contributions to open-source safety tools, publishing research, partnerships and collaborative research initiatives). You may also indicate whether related outputs are publicly shared, shared with partners or within consortia or not publicly disclosed.

Karavela invests in collaborative research on reliable neural decoding, robust evaluation across participants and acquisition settings, and functional MRI data quality. Partnerships with academic researchers and hospitals support scientific validation and responsible data collection. Findings are shared with research partners, with scientific publications and dataset releases planned to support reproducibility and external evaluation.

29. Does your organization pursue research or investment related to socio-economic or environmental considerations in the development and deployment of AI?

- Not applicable

Any further comments and implementation documentation.

No answer provided

Section 7 – Advancing Human and Global Interests

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

No

31. Does your organization pursue AI projects that aim to deliver societal or public benefits, including through partnerships with civil society and community groups?

No

Any further comments and implementation documentation.

No answer provided

