OECD AI Transparency Report

Organization: ai21 (IL) **Reporting Period:** Q1 2025 **Published:** March 20, 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?

AI21 has developed a set of tenets that map directly to the OECD values-based AI principles; Inclusive growth, sustainable development and well-being, Human rights and democratic values, including fairness and privacy, Transparency and explainability and Robustness, security and safety. Given that alignment, the risks defined and classified by the OECD AI Principles correspond with the risks defined and classified by AI21.

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

At a high-level, AI21 classifies risks using the OECD values-based AI principles; Inclusive growth, sustainable development and well-being, Human-centered values and fairness, Transparency and explainability and Robustness, security and safety. The final principle of accountability is focused on the role of AI21, as a company and a set of individuals, in taking responsibility for the behavior of the models, including risk classification and mitigation. We submit that this accountability is demonstrated primarily through transparency and engagement with customers, regulators and independent third parties, such as our engagement with OECD and Stanford University’s CRFM/FMTI. These tenets are used to direct model development and deployment
throughout the full lifecycle; in pre-training, instruct-training, red-teaming, filtering and guard railing and into production testing with customers.

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?

AI21 hires external third-party testing firms to provide independent evaluation of our models. Often the testing is also run internally and the results are compared to those from 3rd-parties. This includes both automated testing as well as human testing of our technologies. We often use industry standard benchmarks in addition to our internal tenets. For example, the RealToxicity, ToxiGen, TruthfulQA

benchmarks are used as just one part of our safety testing. In the red-team testing phase of development, thousands of “attack prompts” are created for each of our 60 tenets to challenge the language model and entice it to break the behavioral expectations of the tenet. Multiple rounds of red-teaming are conducted with
human review to bring the model into alignment and mitigate risk. This is just one example in one phase of the AI lifecycle where risks are evaluated and mitigated.

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 both quantitative and qualitative metrics and focus on competitive benchmarks. That said, our core tenets guide our goals and final implementation. Yes, we often publish whitepapers on our website and on Arxiv to share our findings with a diverse set of stakeholders. We have not implemented an incentive program, but rather hire third-party experts for our evaluations and testing.

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?

Yes, AI21 hires external third-party testing firms to provide independent evaluation of our models. Often the testing is also run internally and the results are compared to those from third-parties. This includes both automated testing as well as human testing of our technologies. We often use industry standard benchmarks in addition to our internal tenets.

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?

Yes, we leverage several risk frameworks including NIST and the EU AI Act.

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 share our findings and contribute to several industry forums, including the OECD Expert Group on AI Risk and Accountability.

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?

AI21 performs evaluations throughout the full lifecycle of development and deployment; in pre-training, instruct-training, red-teaming, filtering and guard railing and into production testing with customers.

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

All of our testing is iterative in nature and the results from evaluations are used to improve the model in subsequent training and development phases. Where risks are found to overlap and relate to one another, priority is typically raised and additional testing/iteration/training is performed.

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

AI21 performs evaluations throughout the full lifecycle of development and deployment; in pre-training, instruct-training, red-teaming, filtering and guard railing and into production testing with customers.

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

AI21 has developed a set of tenets that map directly to the OECD values-based AI principles; Inclusive growth, sustainable development and well-being, Human-centered values and fairness, Transparency and explainability and Robustness, security and safety. The principle of accountability is focused on AI21’s, as a company and a set of individuals, role in taking responsibility for the behavior of the models. We submit that this accountability is demonstrated primarily through transparency and engagement with customers, regulators and independent 3rd-parties. Our engagement with OECD, Stanford University’s CRFM/FMTI validates this commitment to accountability. These tenets are used to direct model development and deployment throughout the full lifecycle; in pre-training, instruct-training, red-teaming, filtering and guard railing and into production testing with customers. For example, in the red-team testing phase of development, thousands of “attack prompts” are created for each of our 60

tenets to challenge the language model and entice it to break the behavioral expectations of the tenet. Multiple rounds of red-teaming are conducted with human review to bring the model into alignment and mitigate risk. This is just one example in one phase of the AI lifecycle where risks are evaluated and mitigated.

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

The creation of a training dataset can be viewed as a pipeline consisting of selection, curation, filtering, augmentation and ingestion. This process is iterative and involves both human and machine evaluation in each phase of the pipeline. Employees of AI21 are involved in every phase and third-party organizations are used in the filtering and augmentation phases of the data pipeline and in later testing (e.g. red-teaming) to provide external review and validation. In the training process, we excluded sites with robot files indicating the presence of copyright material and/or PII.

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

As described above, measures to filter and protect privacy and intellectual property are implemented in every stage of development and also enforced through our responsible use guidelines and terms of services that govern how customers are allowed to use our models and their output.

**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?
**

AI21 implements security controls at every stage of the life-cycle of development and deployment; in pre-training, instruct-training, red-teaming, filtering and guard railing and into production testing with customers. We work closely with our hyper-scaler partners to ensure physical and cyber security of our models and model weights. AI21 works closely with our hyper-scalers to ensure physical and cyber security of our models and model weights.

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

AI21 conducts regular audits of cybersecurity, physical security, insider usage and external usage of the models.

Any further comments and for implementation documentation

https://trust.ai21.com/

https://ai21.com/privacy-policy

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.
**

We publish and share research and discoveries of systemic risks/mitigations and that may impact the systems of our competitors, partners and customers. AI21’s roots are in academic research and we continue to publish research in these areas on our website and in public forums like ArXiv and HuggingFace. We recently released a new hybrid SSM-transformer model as open source to advance model architecture. Similarly, we regularly engage with competitors on risk discovery and mitigation.

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 and share research and discoveries of systemic risks/mitigations and that may impact the systems of our competitors, partners and customers. AI21’s roots are in academic research and we continue to publish research in these areas on our website and in public forums like ArXiv and HuggingFace. We recently released a new hybrid SSM-transformer model as open

source to advance model architecture. Similarly, we regularly engage with competitors on risk discovery and mitigation.

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

Yes - see https://www.ai21.com/privacy-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, we publish model cards and technical whitepapers that provide information about data used for training.

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

Yes, we participate in Stanford's FMTI reporting framework and, of course, the OECD reporting framework in addition to the academic and technical publishing we do.

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?

AI21 has developed and implemented a risk-based approach to identifying and mitigating risk throughout the full lifecycle of development and deployment; in pre-training, instruct-training, red-teaming, filtering and guard railing and into production testing with customers.AI21 performs evaluations throughout the full lifecycle of development and deployment; in pre-training, instruct-training, red-teaming, filtering and guard railing and into production testing with customers. These policies and procedures are updated regularly. For example, with each major release of our models and on an as-needed basis when risks are identified and mitigations developed.

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

Yes, our leadership and employees have regular/required training on these topics.

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

Yes - see https://trust.ai21.com

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

Yes, we have a team dedicated to risk and they maintain a log of reported issues.

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

We publish and share research and discoveries of systemic risks/mitigations and that may impact the systems of our competitors, partners and customers. AI21’s roots are in academic research and we continue to publish research in these areas on our website and in public forums like ArXiv and HuggingFace. We recently released a new hybrid SSM-transformer model as open source to advance model architecture. Similarly, we regularly engage with competitors on risk discovery and mitigation. Finally, we are committed to transparency with the industry as evidenced by a top score in Stanford University’s latest Foundation Model Transparency Index.

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?

We publish and share research and discoveries of systemic risks/mitigations and that may impact the systems of our competitors, partners and customers. AI21’s roots are in academic research and we continue to publish research in these areas on our website and in public forums like ArXiv and HuggingFace. We recently released a new hybrid SSM-transformer model as open source to advance model architecture. Similarly, we regularly engage with competitors on risk discovery and mitigation. Finally, we are committed to transparency with the industry as evidenced by a top score in Stanford University’s latest Foundation Model Transparency Index.

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

We publish and share research and discoveries of systemic risks/mitigations and that may impact the systems of our competitors, partners and customers. AI21’s roots are in academic research and we continue to publish research in these areas on our website and in public forums like ArXiv and HuggingFace. We recently released a new hybrid SSM-transformer model as open source to advance model architecture. Similarly, we regularly engage with competitors on risk

discovery and mitigation. Finally, we are committed to transparency with the industry as evidenced by a top score in Stanford University's latest Foundation Model Transparency Index.

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

See https://trust.ai21.com/ for full list. SOC 2 Type II ISO 27001 ISO 27017 ISO 27018

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 terms of use and responsible use guidelines require that customers not mislead users about interactions with our models. In addition several of our tenets and behavioral expectations for our models direct the model to not present itself as a human on many dimensions. https://docs.ai21.com/docs/responsible-use https://studio.ai21.com/terms-of-use

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?

We are developing technology in this area, however, disclosure of the details can degrade it's efficacy.

Any further comments and for implementation documentation

No answer provided

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?

AI21’s roots are in academic research and we continue to publish research in these areas on our website and in public forums like ArXiv and HuggingFace. We recently released a new version of our hybrid SSM-transformer model as open source to advance model architecture. Similarly, we regularly engage with competitors on risk discovery and mitigation.

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

AI21’s roots are in academic research and we continue to publish research in these areas on our website and in public forums like ArXiv and HuggingFace

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?

Yes.

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

AI21 has developed a set of tenets that map directly to the OECD values-based AI principles; Inclusive growth, sustainable development and well-being, Human-centered values and fairness, Transparency and explainability and Robustness, security and safety. The principle of accountability is focused on AI21’s, as a company and a set of individuals, role in taking responsibility for the behavior of the models. We submit that this accountability is demonstrated primarily through transparency and engagement with customers, regulators and independent 3rd-parties. Our engagement with OECD, Stanford University’s CRFM/FMTI validates this commitment to accountability. These tenets are used to direct model development and deployment throughout the full lifecycle; in pre-training, instruct-training, red-teaming, filtering and guard railing and into production testing with customers. For example, in the red-team testing phase of development, thousands of “attack prompts” are created for each of our 60 tenets to challenge the language model and entice it to break the behavioral expectations of the tenet. Multiple rounds of red-teaming are conducted with human review to bring the model into alignment and mitigate risk. This is just one example in one phase of the AI lifecycle where risks are evaluated and mitigated.

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.

AI21 has developed a set of tenets that map directly to the OECD values-based AI principles; Inclusive growth, sustainable development and well-being, Human-centered values and fairness, Transparency and explainability and Robustness, security and safety. The principle of accountability is focused on AI21's, as a company and a set of individuals, role in taking responsibility for the behavior of the models. We submit that this accountability is demonstrated primarily through transparency and engagement with customers, regulators and independent 3rd-parties. Our engagement with OECD, Stanford University's CRFM/FMTI validates this commitment to accountability. These tenets are used to direct model development and deployment throughout the full lifecycle; in pre-training, instruct-training, red-teaming, filtering and guard railing and into production testing with customers. For example, in the red-team testing phase of development, thousands of "attack prompts" are created for each of our 60 tenets to challenge the language model and entice it to break the behavioral expectations of the tenet. Multiple rounds of red-teaming are conducted with human review to bring the model into alignment and mitigate risk. This is just one example in one phase of the AI lifecycle where risks are evaluated and mitigated.

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.

We provide education and training for our enterprise customers as part of our implementation work with them.

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.

AI21 has developed a set of tenets that map directly to the OECD values-based AI principles; Inclusive growth, sustainable development and well-being, Human-centered values and fairness, Transparency and explainability and Robustness, security and safety. These map to the UN SDGs in many respects.

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.

AI21 has developed a set of tenets that map directly to the OECD values-based AI principles; Inclusive growth, sustainable development and well-being, Human-centered values and fairness,

Transparency and explainability and Robustness, security and safety. These map to the UN SDGs in many respects.

Any further comments and for implementation documentation

https://www.ai21.com/research/ai-code-of-conduct/