Multi-stakeholder Consultation

FUTURE-PROOF AI ACT: TRUSTWORTHY GENERAL-PURPOSE AI

1. SUBJECT

The European AI Office is launching a multi-stakeholder consultation on trustworthy general-purpose AI models in the context of the AI Act.

This contributes to the proper application of Regulation (EU) 2024/1689, referred to as AI Act.

2. BACKGROUND

The European AI Office will support the development and use of trustworthy AI, while protecting against AI risks. Recent advancements in AI has given rise to ever more powerful AI. General-purpose AI models that can perform a wide range of distinct tasks and are being integrated in numerous AI systems are becoming too fundamental for the economy and society not to be regulated.

For fair sharing of responsibilities along the AI value chain and in light of systemic risks, the EU AI Act puts in place effective rules and oversight for general-purpose AI model providers.

Obligations for providers of all general-purpose AI models include, according to Article 53 of the AI Act, keeping up-to-date technical documentation and making such documentation available, upon request, to the AI Office and national competent authorities, as well as providing certain information and documentation to downstream providers integrating the model into their AI system. Providers shall also put in place a policy to comply with Union copyright law as well as make publicly available a summary about the content used for model training, according to a template to be provided by the AI Office.

The providers of general-purpose AI models that are released under a free and open-source license and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available should be subject to exceptions as regards the documentation and information requirements imposed on general-purpose AI models, unless they can be considered to present a systemic risk.

Certain AI models may pose systemic risks if they are highly capable or widely used. In such cases, the AI Act provides additional rules in Article 55 of the AI Act. Providers of general-

purpose AI models with systemic risk are required to assess and mitigate these risks. This includes performing model evaluations and conducting adversarial tests, ensuring an adequate level of cybersecurity protection for both the general-purpose AI model and the physical infrastructure of the model, and keeping track of, documenting, and reporting serious incidents and possible corrective measures.

Rules will be detailed in a Code of Practice, according to Article 56 of the AI Act, which will reflect the state of the art and duly take into account a diverse set of perspectives. For this purpose, the AI Office has recently launched a call for expression of interest to participate in the drawing-up of the first Code of Practice. As detailed out in the call, all interested and eligible general-purpose AI model providers, downstream providers and other industry organisations, other stakeholder organisations such as civil society or rightsholders organisations, as well as academia and other independent experts can join the iterative drafting process.

The AI Office will establish four Working Groups within the Plenary to facilitate the drafting:

- **Working Group 1: Transparency and copyright-related rules**
  Detailing out documentation to downstream providers and the AI Office on the basis of Annexes XI and XII to the AI Act, policies to be put in place to comply with Union law on copyright and related rights, and making publicly available a summary about the training content.
- **Working Group 2: Risk identification and assessment measures for systemic risks**
  Detailing the risk taxonomy based on a proposal by the AI Office and identifying and detailing relevant technical risk assessment measures, including model evaluation and adversarial testing.
- **Working Group 3: Risk mitigation measures for systemic risks**
  Identifying and detailing relevant technical risk mitigation measures, including cybersecurity protection for the general-purpose AI model and the physical infrastructure of the model.
- **Working Group 4: Internal risk management and governance for general-purpose AI model providers**
  Identifying and detailing policies and procedures to operationalise risk management in internal governance of general-purpose AI model providers, including keeping track of, documenting, and reporting serious incidents and possible corrective measures.

Rules for general purpose-AI models will apply from 12 months from the date of entry into force of the AI Act on 1 August 2024. In view of enabling providers to demonstrate compliance on time, the first Code of Practice should be ready by 9 months from the date of entry into force of the Regulation.

In parallel, the AI Office will work on the preparation of the template for the summary about the training data and accompanying guidance, which will set the framework and the minimum level of detail to be covered in the summary, that can be further specified in the Code of Practice.

The AI Office will also work in parallel on the development of the template for the summary about the training content and accompanying guidance that should be provided early enough in the process for general-purpose AI model providers to be able to take them into account as a minimum baseline for further details that can be provided for the summary in the Code of Practice.

3. HAVE YOUR SAY IN THE CONSULTATION

This is an opportunity for all stakeholders to have their say on the topics covered by the first Code of Practice, which will detail out rules for general-purpose AI model providers. It will also inform related work of the AI Office, in particular on the template for the summary of the model training data and the accompanying guidance.

The AI Office therefore launches this consultation on trustworthy general-purpose AI models in the context of the AI Act. We invite submissions from all stakeholders with relevant expertise and perspectives, particularly from civil society organisations, rightsholders organisations, academia or other independent experts, industry such as general-purpose AI model providers and downstream providers, and public authorities.

This consultation aims to gather a broad range of input and perspectives.

The consultation consists of targeted questions and allows for submission of additional material relevant to drawing up the Code of Practice. Answers and submissions will form the basis of the first drafting iteration of the Code. From the outset, the Code will be informed by a wide range of perspectives and expertise. In addition, the consultation results will be used by the AI Office for the development of the template and the accompanying guidance.

The consultation is available in English and responses can be submitted via this consultation form over a period of six weeks. Submissions must be completed by Tuesday, 10 September 2024, 18:00 CET. We encourage early submissions.

The questionnaire for this consultation is structured along 3 sections:

1. General-purpose AI models: transparency and copyright
    - A. Information and documentation to providers of AI systems
    - B. Technical documentation to the AI Office and the national competent authorities
    - C. Policy to respect Union copyright law
    - D. Summary about content used for the training of general-purpose AI models
2. General-purpose AI models with systemic risk
    - A. Risk taxonomy
    - B. Risk identification and assessment
    - C. Technical risk mitigation
    - D. Internal risk management and governance for general-purpose AI model providers
3. Reviewing and monitoring the General-Purpose AI Code of Practice

We welcome full or partial replies from all respondents based on their expertise and perspective.

At the end of the questionnaire, you have the option to upload one document to share further information with the AI Office. We provide a template which aligns with the topics covered in the Code of Practice and follows the structure of the Plenary Working Groups.

All contributions to this consultation may be made publicly available. Therefore, please do not share any confidential information in your contribution. For organisations, their organisation details would be published while respondent details can be requested to be anonymised. Individuals can request to have their contribution fully anonymised.

The AI Office will publish a summary of the results of the consultation. Results will be based on aggregated data and respondents will not be directly quoted.

Please allow enough time to submit your application before the deadline to avoid any issues. In case you experience technical problems, which prevent you from submitting your application within the deadline, please take screenshots of the issue and the time it occurred.

In case you face any technical difficulties or would like to ask a question related to expression of interest, please contact CNECT-AIOFFICE-CODES-OF-PRACTICE@ec.europa.eu