Contribution ID: 275a8f4d-3294-411f-bb8a-b933245e489c
Date: 18/09/2024 14:10:15

Multi-stakeholder Consultation FUTURE-PROOF AI ACT: TRUSTWORTHY GENERAL-PURPOSE AI

Fields marked with * are mandatory.

Multi-stakeholder Consultation FUTURE-PROOF AI ACT: TRUSTWORTHY GENERAL-PURPOSE AI

The European AI Office is launching this multi-stakeholder 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 academia, independent experts, industry representatives such as general-purpose AI model providers or downstream providers integrating the general-purpose AI model into their AI system, civil society organisations, rightsholders organisations, and public authorities.

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

Details about the AI Act rules for providers of general-purpose AI models, the Code of Practice, and related work by the AI Office can be found in the background documents available here.

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

early submissions.

In parallel, stakeholders who wish to participate in the entire process of drawing-up the first Code of Practice can express their interest here by Sunday, 25 August 2024, 18:00 CET.

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. Based on the submissions and answers to the targeted questions, a first draft of the Code of Practice will be developed.

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, please contact: CNECT-AIOFFICE-CODES-OF-PRACTICE@ec.europa.eu

**The AI Office has announced an extension of the consultation period for the Code of Practice concerning general-purpose AI models, as part of the ongoing implementation of the AI Act. The new deadline, set for 18 September 2024, replaces the previous 10 September cutoff. This will grant stakeholders overall seven weeks to submit their feedback.*

About you

* 1. Do you represent one or more organisations (e.g., industry organisation or civil society organisation) or act in your personal capacity (e.g., independent expert)?

- (●) Organisation(s)
- ( ) In a personal capacity

*

Please specify the name(s) of the organisation(s):

Actuarial Association of Europe

* First name

Stephanos

* Surname

Hadjistyllis

* E-Mail address (this won't be published)

stephanos@shsactuarial.com

* Is your organisation headquartered in the EU?

- (●) Yes
- ( ) No
- ( ) Other (e.g. multiple organisations)

* EU member states

- ( ) AT - Austria
- (●) BE - Belgium
- ( ) BG - Bulgaria
- ( ) HR - Croatia
- ( ) CY - Cyprus
- ( ) CZ - Czechia
- ( ) DK - Denmark
- ( ) EE - Estonia
- ( ) FI - Finland
- ( ) FR - France
- ( ) DE - Germany
- ( ) EL - Greece
- ( ) HU - Hungary
- ( ) IE - Ireland
- ( ) IT - Italy
- ( ) LV - Latvia
- ( ) LT - Lithuania
- ( ) LU - Luxembourg
- ( ) MT - Malta
- ( ) NL - Netherlands
- ( ) PL - Poland
- ( ) PT - Portugal
- ( ) RO - Romania

- SK - Slovak Republic
- SI - Slovenia
- ES - Spain
- SE - Sweden

* What is the size of your organisation?
- Micro (1 to 9 employees)
- Small (10 to 49 employees)
- Medium (50 to 249 employees)
- Large (250 or more employees)
- (•) Other (e.g. multiple organisations)

* Please specify

The Actuarial Association of Europe (AAE) represents 38 national actuarial associations in 37 European countries, representing over 30,000 actuaries working in areas including insurance, risk management and pensions. The AAE was established to advise the EU institutions on actuarial matters.
Actuaries in insurance companies and occupational pension providers are responsible for protecting policyholders and beneficiaries by ensuring that the calculations made by the company are appropriate, fair and correct, so that the interests of policyholders and beneficiaries are safeguarded. This responsibility is enshrined in law in most EU countries.

* Which stakeholder category would you consider yourself in?
- Provider of a general-purpose AI model, or acting on behalf of such providers
- Downstream provider of an AI system based on general-purpose AI models, or acting on behalf of such providers
- Other industry organisation, or acting on behalf of such organisations
- Academia
- Civil Society Organisation
- Rightsholder or a collective management organisation (CMO) or an independent management organisation (IME) or the representative of an organisation acting on behalf of rightsholders (other than a CMO or IME)
- Public authority
- (•) Others

* Please specify

Professional Association

* Please briefly describe the activities of your organisation or yourself:

1000 character(s) maximum

The AAE's key objectives are to enhance relations with European institutions; to promote professionalism; and to promote a European community of actuaries.
We play a prominent role in shaping the development of new European legislation touching insurance /pensions, and in the review/refinement of existing legislation, affecting the work of actuaries in traditional areas and in wider fields as actuaries extend their areas of involvement. Additionally, we promote consistent standards of education and professionalism among actuaries in Europe, for example, by prescribing minimum requirements for the education, technical standards, and a code of professional conduct.
Considering the relevant of AI models in actuarial work and the insurance sector, we are responding to this consultation having the actuaries' in mind who may potentially be end users. Whilst actuaries and insurers are not likely to be developers / providers of GPAI products, they may potentially be users of such models.

* **Availability for a follow-up conversation**

We may follow up with you for clarification or further discussion if your submission prompts additional interest.

I agree to be contacted by the AI Office for a follow-up conversation to my submission.

- (●) Yes
- ( ) No

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. Your e-mail address will never be published.

Please select the privacy option that best suits you. Privacy options default based on the type of respondent selected.

* **Contribution publication privacy settings**
If you represent one or more organisations: All contributions to this consultation may be made publicly available. You can choose whether you would like respondent details to be made public or to remain anonymous.

- **Anonymous.** Only organisation details are published: The type of respondent that you responded to this consultation as, the name of the organisation on whose behalf you reply as well as its size, its presence in or outside the EU and your contribution will be published as received. Your name will not be published. Please do not include any personal data in the contribution itself if you want to remain anonymous.
- **Public.** Organisation details and respondent details are published: The type of respondent that you responded to this consultation as, the name of the organisation on whose behalf you reply as well as its size, its presence in or outside the EU and your contribution will be published as received. Your name will also be published.

Privacy statement

☑ I acknowledge the attached privacy statement.

privacy_statement.pdf

Section 1. General-purpose AI models: transparency and copyright-related rules

A. Information and documentation by general-purpose AI model providers to providers of AI systems

Providers of general-purpose AI models have a particular role and responsibility along the AI value chain, as the models they provide may form the basis for a range of downstream systems, often provided by downstream providers that necessitate a good understanding of the models and their capabilities, both to enable the integration of such models into their products, and to fulfil their obligations under the AI Act or other regulations. Therefore, model providers should draw up, keep up-to-date and make available information and documentation to providers of AI systems who intend to integrate the general-purpose AI model into their AI system. Widely adopted documentation practices include model cards and data sheets.

A minimal set of elements of information and documentation by general-purpose AI model providers to providers of AI systems is already set out in AI Act Annex XII.

1. In the **current state of the art**, for which elements of **information and documentation** by general-purpose AI model providers to providers of AI systems do **practices** exist that, in your view, achieve the **above-mentioned purpose**?

From the list below following AI Act Annex XII, please select all relevant elements.
If such practices exist, please provide **links to relevant material** substantiating your reply, such as model cards, data sheets or templates.

A general description of the general-purpose AI model including:

- ☑ **The tasks that the model is intended to perform and the type and nature of AI systems into which it can be integrated;**
- ☑ **The acceptable use policies applicable;**
- ☑ **The date of release and methods of distribution;**
- ☑ **How the model interacts, or can be used to interact, with hardware or software that is not part of the model itself, where applicable;**
- ☑ **The versions of relevant software related to the use of the general-purpose AI model, where applicable;**
- ☑ **The architecture and number of parameters;**
- ☑ **The modality (e.g., text, image) and format of inputs and outputs;**
- ☑ **The licence for the model.**

A description of the elements of the model and of the process for its development, including:

- ☑ **The technical means (e.g., instructions for use, infrastructure, tools) required for the general-purpose AI model to be integrated into AI systems;**
- ☑ **The modality (e.g., text, image, etc.) and format of the inputs and outputs and their maximum size (e.g., context window length, etc.);**
- ☑ **Information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies.**

Alternatively:

- ☐ No practices for any of the listed elements exist that achieve the above-mentioned purpose.

- [ ] I don't know

Links to relevant material

In addition to indicating the model's intended use, the documentation should also clearly state applications where the model should not be used.

EIOPA's CEG on Digital Ethics in Insurance recommends extending the same data quality and governance standards to third party vendors (section IX https://www.eiopa.europa.eu/document/download/30f4502b-3fe9-4fad-b2a3-aa66ea41e863_en?filename=Artificial%20intelligence%20governance%20principles.pdf)

Similarly, the European Actuarial Standard of Actuarial Practice 1 (ESAP1) (https://www.institutdesactuaires.com/global/gene/link.php?doc_id=17919&fg=1) provides relevant recommendations for data quality (see section 2.5)

The description of elements should be tailored to the intended audience. Good practice would be to include a high-level description in plain English for a non-technical audience, followed by an Annex with more technical details for area experts. ESAP1's section 3 Communication, could be relevant in this regard.

Additionally, some examples of the documentation provided can be found on the OpenAI website:
https://platform.openai.com/docs/models/continuous-model-upgrades
https://platform.openai.com/docs/assistants/overview
https://platform.openai.com/docs/guides/text-generation
https://platform.openai.com/docs/guides/vision
https://platform.openai.com/docs/guides/structured-outputs
https://openai.com/policies/

Technical paper that usually comes together with the release of a new LLM: https://ai.meta.com/research/publications/the-llama-3-herd-of-models/
https://huggingface.co/

2. Beyond the minimal set of elements listed in the previous question, are there **other elements** that should be included in **information and documentation** by general-purpose AI model providers to providers of AI systems to achieve the above-mentioned purpose?

- (•) Yes
- ( ) No
- ( ) I don't know

Please specify

700 character(s) maximum

Providers should indicate known or suspected instances / conditions where their models are known to be prone to make errors or where their outputs should not be trusted. They should also provide a high-level summary of data sources, data transformations, and whether any such transformations were to address known biases.

We note that actuaries abide by European Standards of Actuarial Practice (ESAPs) which are model standards of practice for actuarial work (link provided below). Documentation provided for GPAI models should be sufficient so that actuaries can fulfill their technical and professional standards. Please refer to ESAP1 in the link below.

Links to relevant material

https://actuary.eu/wp-content/uploads/2019/10/ESAP1-Revised-11-10-2019-FINAL.pdf

B. Technical documentation by general-purpose AI model providers to the AI Office and the national competent authorities

In addition to the provision of information on the general-purpose AI model for its usage by the downstream providers, technical documentation should be prepared and kept up to date by the general-purpose AI model provider for the purpose of making it available, upon request, to the AI Office and the national competent authorities.

A minimal set of elements of such technical documentation of the general-purpose AI model to be made available by providers, upon request, to the AI Office and the national competent authorities is already set out in AI Act Annex XI.

3. In the **current state of the art**, for which elements of **documentation** by general-purpose AI model providers do practices exist that, in your view, provide a **necessary level of information for the above-mentioned purpose**?

From the list below following AI Act Annex XI, please select all relevant elements.
If such practices exist, please provide **links to relevant material** substantiating your reply, such as model cards, data sheets or templates.

A general description of the general-purpose AI model including:

- ☑ **The tasks that the model is intended to perform and the type and nature of AI systems into which it can be integrated**;

- [x] **The acceptable use policies applicable;**
- [x] **The date of release and methods of distribution;**
- [x] **The architecture and number of parameters;**
- [x] **The modality (e.g., text, image) and format of inputs and outputs;**
- [x] **The licence.**

A description of the elements of the model, and relevant information of the process for the development, including:

- [x] **The technical means (e.g., instructions for use, infrastructure, tools) required for the general-purpose AI model to be integrated into AI systems;**
- [x] **The design specifications of the model and training process**, including training methodologies and techniques, the key design choices including the rationale and assumptions made; what the model is designed to optimise for and the relevance of the different parameters, as applicable;
- [x] **Information on the data used for training, testing and validation**, where applicable, including the type and provenance of data and curation methodologies (e.g. cleaning, filtering etc), the number of data points, their scope and main characteristics; how the data was obtained and selected as well as all other measures to detect the unsuitability of data sources and methods to detect identifiable biases, where applicable;
- [ ] **the computational resources used to train the model** (e.g. number of floating point operations), training time, and other relevant details related to the training;
- [ ] **known or estimated energy consumption of the model.**

Additional information to be provided by providers of general-purpose AI models with systemic risk:

- [ ] **A detailed description of the evaluation strategies**, including evaluation results, on the basis of available public evaluation protocols and tools or otherwise of other evaluation methodologies. Evaluation strategies shall include evaluation criteria, metrics and the methodology on the identification of limitations;
- [ ] **Where applicable, a detailed description of the measures put in place for the purpose of conducting internal and/or external adversarial testing** (e. g., red teaming), model adaptations, including alignment and fine-tuning;

- [ ] **Where applicable, a detailed description of the system architecture** explaining how software components build or feed into each other and integrate into the overall processing;

Alternatively:

- [ ] No practices for any of the listed elements exist that achieve the above-mentioned purpose.
- [ ] I don't know

Links to relevant material

EIOPA's CEG on Digital Ethics in Insurance recommends that adequate governance is applied throughout the entire lifecycle of the model therefore the AI Office should receive sufficient information to assess the quality of this governance (https://www.eiopa.europa.eu/document/download/30f4502b-3fe9-4fad-b2a3-aa66ea41e863_en?filename=Artificial%20intelligence%20governance%20principles.pdf)]

An Huggingface model is a relevant one to consider: https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

4. Beyond the minimal set of elements listed in the previous question, are there **oth er elements** that should, in your view, be included in **technical documentation** by general-purpose AI model providers **to the AI Office** and the national competent authorities?

- (•) Yes
- ( ) No
- ( ) I don't know

Please specify

700 character(s) maximum

Providers should indicate known or suspected instances / conditions where their models are known to be prone to make errors or where their outputs should not be trusted. They should also provide a high-level summary of data sources, data transformations, and whether any such transformations were to address known biases.
Documentation setting out the model providers' ongoing risk management procedures, to ensure that the model is fit for use, and meets ethical standards, not just at the outset, but as it evolves over time.
Emergent abilities in GENAI models are unexpected skills that appear as model size grows. Providers should quantify these risks and detail measures taken to control them.

Links to relevant material

C. Policy to respect Union copyright law

The AI Act requires providers of general-purpose AI models to put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790.

5. What are, in your view, the main **elements that need to be included in the policy** that providers of general-purpose AI models have to put in place to **comply with Union law on copyright** and related rights, as required by the AI Act?

Please select all relevant options from the list of options suggested below. If selected, please elaborate further on the content of the measures and provide links to any good practices you are aware of.

- ☐ Allocation of responsibility within the organisation for the implementation and monitoring of compliance with the policy and the measures therein;
- ☐ Measures to identify and comply with the rights reservation from the text and data mining exception pursuant to Article 4(3) of Directive (EU) 2019/790;
- ☐ Measures to obtain the authorisation from right holders, where applicable;
- ☐ Measures to detect and remove collected copyright protected content for which rights reservation from the text and data mining exception has been expressed pursuant to Article 4(3) of Directive (EU) 2019/790;
- ☐ Measures to prevent the generation, in the outputs of the model, of copyright infringing content;
- ☐ Means for contact with rightsholders;
- ☐ Measures for complaint handling from rightsholders;
- ☐ Other
- ☑ I don't know

Your comments

700 character(s) maximum

Our expertise in copyright law is fairly limited and we have therefore abstained from answering this question.

Links to relevant material

6. How can, in your view, the policy to be put in place by providers of general-purpose AI models to comply with Union copyright law ensure that providers of those models comply with the **existing solutions for the expression of the text and data mining rights reservation**, pursuant to Article 4(3) of Directive (EU) 2019 /790?

Please explain how this can be achieved and specify from the list below the state-of-the-art technologies you are aware of to identify and comply with the right reservations expressed by rightsholders, providing further information and examples.

- [ ] Technologies/tools that identify right reservations at the website/domain level
- [ ] Technologies/tools that identify right reservations at work level
- [ ] Technologies/tools that aggregate the expression of right reservations
- [ ] Other
- [x] I don't know

Your comments

700 character(s) maximum

Our expertise in copyright law is fairly limited and we have therefore abstained from answering this question.

Links to relevant material

D. Summary about content used for the training of general-purpose AI models

The AI Act requires providers to draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office. While due account should be taken of the need to protect trade secrets and confidential business information, the summary is to be generally comprehensive in its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law. The template that should be drafted by the AI Office for the sufficiently detailed summary should be simple, effective, and allow providers to provide the required summary in narrative form.

7. What are in your view the **categories of information** sources that should be presented in the summary to ensure that it comprehensively describes the main sources of data used for the training of the general-purpose AI model?

From the list below, please select all options that you consider relevant.

- [x] Public/ open data repositories
- [x] Content/data publicly available online (e.g. scraped from the internet)
- [x] Proprietary data generated by the provider
- [x] User-generated data obtained through the services or products provided by the provider
- [x] Copyright protected content licensed by rightsholders
- [x] Other data/content or data sets acquired from third parties (e.g. licensed proprietary databases, data acquired from datahubs, public interest institutions such as libraries etc.)
- [x] Synthetically generated data
- [ ] Other
- [ ] I don't know

If selected, please **specify the level of granularity/detail for each of the selected options**, keeping in mind that AI Act requires the summary to be comprehensive instead of technically detailed and provided in a narrative form to facilitate parties with legitimate interests, including rightsholders, to exercise and enforce their rights under Union law, while taking due account of the need to protect providers' trade secrets and confidential business information. If additional categories should be considered, please specify them and the level of granularity /detail. You can motivate your choice and provide links to any good practices.

700 character(s) maximum

The level of granularity and detail for the selected options and its relevant importance may vary depending on the stakeholder group in question. However, we note that to satisfy the requirements of ESAP1, an actuary needs to know what data/information sources were used. For example, an actuary needs to know there is no inherent bias resulting from the use of different sources, and that there is no danger of inferring causality from correlation.

There should be an indication of the proportions each data source represents in the total, as well as high-level description of data transformations, particularly those carried out to address known biases (or missing values) in the source data.

Links to relevant material

https://actuary.eu/wp-content/uploads/2019/10/ESAP1-Revised-11-10-2019-FINAL.pdf

Please refer to the relevant sections on data.

8. In your view, should the summary include one or more of the following **characteristics/information about the data used for the training**/of the general-purpose AI model in order to facilitate parties with legitimate interests, including copyright holders, to enforce their rights under Union law?

Please select all relevant options from the list of options suggested below. If selected, please explain your choice and provide links to any good practices.

- [x] Modalities / type of data (text, images, videos, music, etc);
- [x] Nature of the data (personal, non-personal or mixed);
- [x] Time of acquisition/collection of the data;
- [x] Data range of the data (e.g. time span), including date cutoffs
- [x] In case of data scraped from the internet, information about the crawlers used;
- [x] Information about diversity of the data (for example linguistic, geographical, demographic diversity);
- [x] Percentage of each of the main data sources to the overall training/fine-tuning;
- [x] Legal basis for the processing under Union copyright law and data protection law, as applicable;
- [x] Measures taken to address risks to parties with legitimate interests (e.g. measures to identify and respect opt-out from the text and data mining exception, respect data protection and address privacy risks, bias, generation of illegal or harmful content;
- [ ] Other
- [ ] I don’t know

Your comments

700 character(s) maximum

Link to relevant material

9. Considering the purpose of the summary to provide **meaningful information to facilitate the exercise of the rights** of parties with legitimate interests under Union law, while taking due account of the need to respect **business confidentiality and trade secrets** of providers, what **types of information** in your view are **justified not to be disclosed** in the summary as being not necessary or disproportionate for its purpose described above?

700 character(s) maximum

Our expertise in this topic is fairly limited as it is of legal nature and we have therefore abstained from answering this question.

Section 2. General-purpose AI models with systemic risk: risk taxonomy, assessment and mitigation

A. Risk taxonomy

Some general-purpose AI models could pose systemic risks, which should be understood to increase with model capabilities and model reach and can arise along the entire lifecycle of the model.

‘Systemic risks’ refer to risks that are specific to the high-impact capabilities of general-purpose AI models (matching or exceeding the capabilities of the most advanced general-purpose AI models); have a significant impact on the Union market due to their reach; or are due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or society as a whole, that can be propagated at scale across the value chain (AI Act Article 3(65)).

Systemic risks are influenced by conditions of misuse, model reliability, model fairness and model security, the level of autonomy of the model, its access to tools, novel or combined modalities, release and distribution strategies, the potential to remove guardrails and other factors.

The Code of Practice should help to establish a risk taxonomy of the type and nature of the systemic risks at Union level, including their sources. The Code should take into account international approaches.

10. Do you consider the following list of **systemic risks** based on AI Act Recital 110 and international approaches to be comprehensive to inform a taxonomy of

systemic risks from general-purpose AI models? If additional risks should be considered in your view, please specify.

Systemic risk from model malfunctions

- **Harmful bias and discrimination:** The ways in which models can give rise to harmful bias and discrimination with risks to individuals, communities or societies.
- **Misinformation and harming privacy:** The dissemination of illegal or false content and facilitation of harming privacy with threats to democratic values and human rights.
- **Major accidents:** Risks in relation to major accidents and disruptions of critical sectors, that a particular event could lead to a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community.
- **Loss of control:** Unintended issues of control relating to alignment with human intent, the effects of interaction and tool use, including for example the capacity to control physical systems, 'self-replicating' or training other models.

Systemic risk from malicious use

- **Disinformation:** The facilitation of disinformation and manipulation of public opinion with threats to democratic values and human rights.
- **Chemical, biological, radiological, and nuclear risks:** Dual-use science risks related to ways in which barriers to entry can be lowered, including for weapons development, design acquisition, or use.
- **Cyber offence:** Risks related to offensive cyber capabilities such as the ways in which vulnerability discovery, exploitation, or operational use can be enabled.

Other systemic risks, with reasonably foreseeable negative effects on

- **public health**
- **safety**
- **democratic processes**
- **public and economic security**

- **fundamental rights**
- **the society as a whole.**

☐ Yes, this list of systemic risks is comprehensive.
☑ Further or more specific systemic risks should be considered.
☐ I don't know

Please specify

700 character(s) maximum

Financial risks to individuals and businesses should also be considered. Similarly, risks to the stability of financial markets such as those arising from algorithmic trading, automated pricing and underwriting etc. are significant. Collaboration with the ESAs (EIOPA, EBA, ESMA) would be beneficial in this regard.

Systemic risk (similar to the 2008 financial crisis) has crucial manifestations in the financial services and we believe it should be addressed thoroughly. The AI Office could work with ECB, ESRB, and ESAs (including EIOPA) to address systemic risk in the financial sector in more detail.

11. What are in your view **sources of systemic risks** that may stem from the development, the placing on the market, or the use of general-purpose AI models? Systemic risks should be understood to increase with model capabilities and model reach.

Please select all relevant elements from the list. If additional sources should be considered, please specify. You can also provide details on any of the sources or other considerations.

☑ **Level of autonomy of the model**: The degree to which a general-purpose AI model has the capability to autonomously interact with the world, plan ahead, and pursue goals.
☑ **Adaptability to learn new, distinct tasks:** The capability of a model to independently acquire skills for different types of tasks.
☑ **Access to tools:** A model gaining access to tools, such as databases or web browsers, and other affordances in its environment.
☐ **Novel or combined modalities:** Modalities a model can process as input and generate as output, such as text, images, video, audio or robotic actions.
☑ **Release and distribution strategies:** The way a model is released, such as under free and open-source license, or otherwise made available on the market.
☑

Potential to remove guardrails: The ability to bypass or disable pre-defined safety constraints or boundaries set up to ensure a model operates within desired parameters and avoids unintended or harmful outcomes.

- [ ] **Amount of computation used for training the model:** Cumulative amount of computation ('compute') used for model training measured in floating point operations as one of the relevant approximations for model capabilities.
- [x] **Data set used for training the model:** Quality or size of the data set used for training the model as a factor influencing model capabilities.
- [x] **Other**
- [ ] **I don't know**

Please specify

700 character(s) maximum

Concentration risk: A lack of competition in a specific area could lead to most participants in a given market to rely on a single provider or AI model. This could have an impact on market stability, exacerbate the harm caused by crystalising risks such as discrimination etc.

Your comments

700 character(s) maximum

B. Risk identification and assessment measures

In light of potential systemic risks, the AI Act puts in place effective rules and oversight. Providers of general-purpose AI models with systemic risks should continuously assess and mitigate systemic risks.

The Code of Practice should be focused on specific risk assessment measures for general-purpose AI models with systemic risk. Following the risk taxonomy, **appropriate measures could be applied to assess different systemic risks, tailored to each specific type and nature of risk,** including their sources.

In addition to further risk assessment measures which will be detailed out in the Code of Practice, the AI Act requires providers to perform the necessary model evaluations, in particular prior to its first placing on the market, including conducting and documenting adversarial testing of the model, also, as appropriate, through internal or independent external testing.

The following concerns technical risk assessment measures, including model evaluation and adversarial testing. This is in line with the focus of the Code of Practice Working Group 2 “Risk identification and assessment measures for systemic risks”.

12. How can the effective implementation of **risk assessment measures reflect differences in size and capacity** between various providers such as SMEs and start-ups?

700 character(s) maximum

Measures should be proportionate to the potential for harm, regardless of the size and capacity of the provider. This would ensure a level playing field without perverse incentives driving providers to remain small to face less stringent risk management requirements. It would also provide better protection for society. The potential for harm should be considered based on the probability of risk materializing, and the damaged caused if it happens. The model’s reach in the market and the number of individuals potentially affected could also be taken in consideration alongside the nature of the harm. EIOPA’s Guidelines on Trustworthy AI in Insurance can be a useful reference.

13. In the **current state of the art**, which specific **risk assessment measures** should, in your view, general-purpose AI model providers take to effectively assess systemic risks along the entire model lifecycle, *in addition* to evaluation and testing?

Please **indicate to what extent you agree** that providers should take the risk assessment measures from the list. You can add additional measures and provide details on any of the measures, such as what is required for measures to be effective in practice.

Potential risk assessment measures	Strongly agree	Somewhat agree	Neither agree nor disagree	Disagree	I don’t know
Determining risk thresholds and risk tolerance, incl. acceptable levels of risks and capabilities for model development and deployment, and respective quantification of risk severity and probability	◉	○	○	○	○
Forecasting model capabilities and risks before and during model development	○	◉	○	○	○
Continuous monitoring for emergence of risks, including data from users, relevant stakeholders, incident databases or similar	◉	○	○	○	○

Determining effectiveness of risk mitigation measures	◉	○	○	○	○
Safety cases to demonstrate that the model does not exceed maximum risk thresholds	○	○	◉	○	○
Aggregate risk assessment before model development	◉	○	○	○	○
Aggregate risk assessment before model deployment	◉	○	○	○	○
Aggregate risk assessment along the entire model lifecycle	◉	○	○	○	○
Third-party involvement in risk assessment, for example, related to inspections of training data, models or internal governance	◉	○	○	○	○

And/or:

○ **Other**

If table is not submitted

○ I don't know

Your comments

700 character(s) maximum

We would suggest that, although the actions above are sensible, the risks to the financial sector should be addressed more specifically (with aid from ECB, ESRB, and ESAs), noting that to effectively assess systemic risks of GPAI models, one has to consider the the specific “downstream” use cases.

Financial services firms should integrate AI risk management within their existing processes as required by BASEL or Solvency II, including processes such as Risk Appetite Statements, ORSA, risk registers, etc. EIOPA’s guidelines on systems of governance could be of relevant for some items, in particular the introduction, guidelines 1 to 8, 17 to 19, 21, 53, and 63).

14. Please provide **links to relevant material** on state-of-the-art risk assessment measures, such as model cards, data sheets, templates or other publications.

15. In the **current state of the art**, which specific practices related to **model evaluations** should, in your view, general-purpose AI model providers take with a view to identifying and mitigating systemic risks?

Model evaluations can include various techniques, such as benchmarks and automated tests, red teaming and adversarial testing including stress testing and boundary testing, white-box evaluations with model explanation and interpretability techniques, and sociotechnical evaluations like field testing, user studies or uplift studies.

Please **indicate to what extent you agree** that providers should implement the practice from the list. You can add additional practices and provide details on any of the practices. You can also indicate which model evaluation techniques listed above or which other techniques can reliably assess which specific systemic risks.

Potential evaluation practices	Strongly agree	Somewhat agree	Neither agree nor disagree	Disagree	I don't know
Performing evaluations at several checkpoints throughout the model lifecycle, in particular **during development and prior to internal deployment**	○	◉	○	○	○
Performing evaluations at several checkpoints throughout the model lifecycle, in particular **when the model risk profile changes** such as with access to tools or with different release strategies	◉	○	○	○	○
Ensuring evaluations inform model deployment in real-world conditions	◉	○	○	○	○
Ensuring evaluations provide insights into the degree to which a model introduces or exacerbates risks	◉	○	○	○	○
Using non-public model evaluations, as appropriate	○	○	◉	○	○
Involve independent external evaluators, including with appropriate levels of access to the model and related information	○	○	◉	○	○
Involve affected persons, to understand effects of human interactions with a particular model over time	○	◉	○	○	○

Documenting evaluation strategies and results	◉	○	○	○	○
Reporting evaluation strategies and results **publicly**, as appropriate	○	○	◉	○	○
Reporting evaluation strategies and results to **selected authorities and administrative bodies**, as appropriate, including sensitive evaluation results	○	◉	○	○	○
Continuously evaluate and improve evaluation strategies based on information from risk assessment and mitigation measures, including from incidents and near-misses	○	◉	○	○	○

And/or:

○ **Other**

It table is not submitted

○ I don't know

Your comments

700 character(s) maximum

16. Please provide **links to relevant material** on state-of-the-art model evaluation practices, such as model cards, data sheets, templates or other publications.

https://actuary.eu/wp-content/uploads/2019/10/ESAP1-Revised-11-10-2019-FINAL.pdf

We would like to once again refer to European Standards of Actuarial Practice and in particular ESAP1 with regards to model governance, data and assumptions.

17. What are the **greatest challenges** that a general-purpose AI model provider could face in implementing risk assessment measures, including model evaluations?

700 character(s) maximum

The unknown unknowns pose a significant challenge. Risk assessment measures can be used with known or anticipated risks (known knowns, and known unknowns). However, it is very difficult to effectively manage risks that providers are not even aware they exist. For example, providers may not assess for potential discrimination against individuals of some protected characteristics if they never come to their mind such as autistic individuals, members of minority religions, etc.

C. Technical risk mitigation

Codes of Practice should also be focused on specific risk mitigation measures for general-purpose AI models with systemic risk. Following the risk taxonomy, **appropriate measures could be applied to mitigate different systemic risks, tailored to each specific type and nature of risk**, including their sources.

The following concerns technical risk mitigation measures, including cybersecurity protection for the general-purpose AI model and the physical infrastructure of the model. Measures can relate to model design, development or deployment.

This is in line with the focus of the Code of Practice Working Group 3 “Risk mitigation measures for systemic risks”.

18. How can the effective implementation of **technical risk mitigation measures reflect differences in size and capacity** between various providers such as SMEs and start-ups?

700 character(s) maximum

As indicated before, risk mitigation measures should be commensurate to the harm the risk may cause, regardless of the size and capacity of the provider. Small providers should ensure that they do not overstretch their capabilities, and that their risk mitigation practices grow at the same time as their market reach.

https://www.eiopa.europa.eu/document/download/30f4502b-3fe9-4fad-b2a3-aa66ea41e863_en?filename=Artificial%20intelligence%20governance%20principles.pdf

19. In the **current state of the art**, which specific **technical risk mitigation measures** should, in your view, general-purpose AI model providers take to effectively mitigate systemic risks along the entire model lifecycle, *in addition* to cybersecurity protection?
Please **indicate to what extent you agree** that providers should take the measures from the list. You can add additional measures and provide details on any of the measures, such as what is required for measures to be effective in practice.

Potential technical risk assessment measures	Strongly agree	Somewhat agree	Neither agree nor disagree	Disagree	I don’t know

Data governance such as data selection, cleaning, quality control	◉	○	○	○	○
Model design and development to achieve an appropriate level of trustworthiness characteristics such as model reliability, fairness or security	◉	○	○	○	○
Fine-tuning for trustworthiness and alignment such as through Reinforcement Learning from Human Feedback (RLHF) or Constitutional AI	○	○	○	○	◉
Unlearning techniques such as to remove specific harmful capabilities from models	◉	○	○	○	○
Technical deployment guardrails, such as content and other filters, capability restrictions, fine-tuning restrictions or monitoring-based restrictions in case of misuse by users	◉	○	○	○	○
Mitigation measures relating to the model architecture, components, access to tools or model autonomy	○	◉	○	○	○
Detection, labelling and other measures related to AI-generated or manipulated content	○	◉	○	○	○
Regular model updates, including security updates	◉	○	○	○	○
Measuring model performance on an ongoing basis	○	◉	○	○	○
Identification and mitigation of model misuse	◉	○	○	○	○
Access control to tools and levels of model autonomy	◉	○	○	○	○

And/or:

○ **Other**

If table is not submitted

○ I don't know

Your comments

700 character(s) maximum

20. Please provide **links to relevant material** on state-of-the-art technical risk mitigation practices, such as model cards, data sheets, templates or other publications.

We would like to refer to ESAP1 noted previously. The sections on model governance, process management, and peer review are quite relevant to consider.

21. What are the **greatest challenges** that a general-purpose AI provider could face in implementing technical risk mitigation measures?

700 character(s) maximum

Overconfidence in the effectiveness of such mitigation techniques. There should be regular monitoring that these mitigation measures work as intended. In addition, it is common to underestimate the severity of a risk and so stress testing of risk mitigation techniques should be carried out using realistic assumptions and scenarios, subject to challenge as appropriate. Ethics also bring an additional risk, where business goals influence the identification or resolution of ethical issues related to the use of data, model performance, intended uses, among other things.

D. Internal risk management and governance for general-purpose AI model providers

The following concerns 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.

This is in line with the focus of the Code of Practice Working Group 4 “Internal risk management and governance for general-purpose AI model providers”.

22. How can the effective implementation of **internal risk management and governance measures reflect differences in size and capacity** between various providers such as SMEs and start-ups?

700 character(s) maximum

Governance and risk management practices should be aligned with the intended use of AI models, and proportional to their potential for harm. SMEs and start-ups with limited capabilities may want to set boundaries to the use of their models to limit the governance and risk management practices required in some cases.

Links to relevant material

23. In the **current state of the art**, which specific **internal risk management and governance measures** should, in your view, general-purpose AI model providers take to effectively mitigate systemic risks along the entire model lifecycle, *in addition* to serious incident reporting?

Please **indicate to what extent you agree** that providers should take the measures from the list. You can add additional measures and provide details on any of the measures, such as what is required for measures to be effective in practice.

Potential internal risk management and governance measures	Strongly agree	Somewhat agree	Neither agree nor disagree	Disagree	I don't know
Risk management framework across the model lifecycle	◉	○	○	○	○
Internal independent oversight functions in a transparent governance structure, such as related to risks and ethics	◉	○	○	○	○
Traceability in relation to datasets, processes, and decisions made during model development	◉	○	○	○	○
Ensuring that staff are familiar with their duties and the organisation's risk management practices	○	◉	○	○	○
Responsible scaling policies	○	◉	○	○	○
Acceptable use policies	○	◉	○	○	○
Whistleblower protections	○	◉	○	○	○
Internal resource allocation towards risk assessment and mitigation measures as well as research to mitigate systemic risks	○	◉	○	○	○
Robust security controls including physical security, cyber security and information security	○	◉	○	○	○
External accountability measures such as third-party audits, model or aggregated data access for researchers	○	○	◉	○	○

Other collaborations and involvements of a diverse set of stakeholders, including impacted communities	○	◉	○	○	○
Responsible release practices including staged release, particularly before open-sourcing a model with systemic risk	○	◉	○	○	○
Transparency reports such as model cards, system cards or data sheets	○	◉	○	○	○
Human oversight mechanisms	○	◉	○	○	○
Know-your-customer practices	○	◉	○	○	○
Logging, reporting and follow-up of near-misses along the lifecycle	○	◉	○	○	○
Measures to mitigate and remediate deployment issues and vulnerabilities	○	◉	○	○	○
Complaints handling and redress mechanisms, such as bug bounty programs	○	◉	○	○	○
Mandatory model updating policies and limit on maximum model availability	○	◉	○	○	○
Third-party and user discovery mechanisms and reporting related to deployment issues and vulnerabilities	○	◉	○	○	○

And/or:

○ **Other**

If table is not submitted

○ I don't know

Your comments

700 character(s) maximum

Due to regulatory requirements in financial services, providers might have to be subject to increased requirements when servicing banks and insurers, including third party external audits.

24. Please provide **links to relevant material** on state-of-the-art governance risk mitigation practices, such as model cards, data sheets, templates or other publications.

25. What are the **greatest challenges** that a general-purpose AI provider could face in implementing governance risk mitigation measures?

700 character(s) maximum

One key challenge is ensuring that those with access to the model have sufficient understanding of how it works, and have been giving appropriate technical and ethical training, so as to be able to identify risks with such complex models. Maintaining these high standards aa models rapidly evolve over time will be a challenge.

General comments:
- The code should be kept to 10 pages maximum to keep it efficient and usable.
- The governance section should be first, not in the second part. And it should focus on responsibility, not reporting.
- With responsibility, transparency of what the model does is the best protection, whilst costs/benefits estimates must always be included.

Section 3. Reviewing and monitoring of the General-Purpose AI Code of Practice

The process of drawing-up the first Code of Practice will start immediately after the AI Act enters into force and will last for 9 months, in view of enabling providers of general-purpose AI models to demonstrate compliance on time. The AI Office shall aim to ensure that the Code of Practice clearly sets out their specific objectives and contains commitments or measures, including key performance indicators as appropriate, to ensure the achievement of those objectives.

The AI Office shall aim to ensure that participants to the Code of Practice report regularly to the AI Office on the implementation of the commitments and the measures taken and their outcomes, including as measured against the key performance indicators as appropriate. Key performance indicators and reporting commitments shall reflect differences in size and capacity between various participants. The AI Office and the Board shall regularly monitor and evaluate the achievement of the objectives of the Code of Practice by the participants and their contribution to the proper application of this Regulation.

The AI Office shall, as appropriate, encourage and facilitate the review and adaptation of the Code of Practice.

26. What are examples of **key performance indicators** which are, in your view, effective to measure the compliance of participants with the objectives and measures which will be established by the Code of Practice?

700 character(s) maximum

While KPIs can be useful, there is the risk of providers focus exclusively on the KPIs and lose sight of the bigger picture. Optimization of KPIs while other risks are ignored would be a danger. It could turn risk management into a tick box exercise where nothing receives due attention other than prescribed KPIs. The best approach would require principles or principles in addition to specific requirements such as KPIs.

In terms of KPIs, the number of participants meeting the expected standard of compliance would be an example. Another one is limiting the number of “significant incidents” of AI risk propagation, akin to data breaches in data protection legislation.

Links to relevant material

27. How can **key performance indicators and reporting commitments** for providers **reflect differences in size and capacity** between various providers such as SMEs and start-ups?

700 character(s) maximum

As discussed previously, the use of KPIs when applied to SMEs and Start-ups should follow a risk-based approach and take into consideration the probability of incidents and expected severity (i.e., impact) that the providers may have.

Links to relevant material

28. Which aspects should inform the timing of **review and adaptation of the content of the Code of Practice** for general-purpose AI models in order to ensure that the **state of the art** is reflected? This does not necessarily imply a complete review, but can only involve pertinent parts.

Please rank all relevant aspects from the following list from 1 to 4 (1 being the most important). You can add additional aspects and provide details on any of the aspects or other considerations under "Specify".

	Rank 1	Rank 2	Rank 3	Rank 4

Pre-planned intervals to assess the need for revision: Assessments of whether the content of the Code of Practice for general-purpose AI models needs to be revised or adapted should be pre-planned for specific time intervals.	○	○	◉	○
Alerts by independent experts or other stakeholders: Alerts by selected independent experts, such as by the Scientific Panel which will be set up in the AI Act governance structure, or by other stakeholders such as downstream providers, academia or civil society should inform a revision of the content of the Code of Practice.	○	◉	○	○
Monitoring and foresight: Independent monitoring and foresight related to the AI ecosystem, technological and market developments, emergence of systemic risks and any other relevant trends, such as related to sources of risks like model autonomy, should inform a revision of the content of the Code of Practice	◉	○	○	○
Other	○	○	○	○

Specify for "**Other**"

If ranking is not submitted

○ **I don't know**

Your comments

700 character(s) maximum

We note that the response to this consultation forms our preliminary view as an organisation and may be subject to change as new information becomes available in the practice of AI in general. As such, we may follow up with additional comments as necessary.

Links to relevant material

Option to upload a document for additional information

You have the option to upload one document to share further information with the AI Office. Please download the template that is structured along the topics covered by the Code of Practice Working Groups. Based on the submissions and answers to the targeted questions, a first draft of the Code of Practice will be developed.

Please upload your document in a doc or docx format, instead of pdf or similar.

Template_for_free-text_submissions.docx

Please upload your file(s)

Only files of the type doc,docx are allowed

Thank you

Thank you for participating in the consultation. Please don't forget to click on submit.

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.

All contributions to this consultation may be made publicly available.

Contact

Contact Form