HUGGING FACE

Hugging Face Comments on the EU AI Office's Multi-stakeholder Consultation FUTURE-PROOF AI ACT: TRUSTWORTHY GENERAL-PURPOSE AI

Responses by: Lucie-Aimée Kaffee, Yacine Jernite, Bruna Trevelin

About Hugging Face

Hugging Face is a community-oriented company working to democratise good Machine Learning (ML), and has become the most widely used platform for sharing and collaborating on ML systems. We are an open-source and open-science platform hosting machine learning models and datasets within an infrastructure that supports easily processing and analysing them; conducting novel AI research; and providing educational resources, courses, and tooling to lower the barrier for all backgrounds to contribute to AI. Part of those activities include releasing open-source tools to support other actors developing their own models, as well as collaborating on new state-of-the-art GPAI models such as StarCoder2, the highest-performing fully-open code model to date. Its training dataset, The Stack, exemplifies a working opt-out methodology, and its governance card provides an implementation of the goals of the Code of Practice: https://shorturl.at/lBbAa.

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.

Links to relevant material

Model cards on Hugging Face: https://huggingface.co/docs/hub/en/model-cards
Example of usage of model cards: https://huggingface.co/meta-llama/Meta-Llama-3-8B (includes intended tasks, use policy, release date, version, architecture and number of parameters, modality, license, technical means)

https://huggingface.co/CohereForAI/c4ai-command-r-v01 (includes intended tasks, use policy, release date, version, architecture and number of parameters, modality, license, technical means)
Example usage of the model cards: https://huggingface.co/HuggingFaceM4/idefics2-8b (includes intended tasks, use policy, release date, version, architecture and number of parameters, modality, license, technical means, dataset for training (see below as OBELICS))
Usage policy (RAIL and OpenRAIL):
https://www.licenses.ai/blog/2022/8/18/naming-convention-of-responsible-ai-licenses
https://huggingface.co/spaces/bigscience/license
Use of RAIL: https://huggingface.co/bigscience/bloom
Tool Use With Command R: https://cohere.com/blog/tool-use-with-command-r
Dataset cards on Hugging Face: https://huggingface.co/docs/hub/en/datasets-cards
https://github.com/huggingface/datasets/blob/main/templates/README_guide.md
Example of usage of dataset cards:
https://huggingface.co/datasets/HuggingFaceM4/OBELICS

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?

- [x] Yes
- [ ] No
- [ ] I don't know

Links to relevant material

Where applicable, ethical charters and other considerations made in the creation of the model: https://bigscience.huggingface.co/blog/bigscience-ethical-charter
https://huggingface.co/blog/ethical-charter-multimodal
Governance cards: https://arxiv.org/abs/2312.03872
https://huggingface.co/datasets/bigcode/governance-card

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

A description of the elements of the model, and relevant information of the process for the 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 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;
- ☑ 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;

Links to relevant material

Example of usage of model cards: https://huggingface.co/meta-llama/Meta-Llama-3-8B (includes intended tasks, use policy, release date, version, architecture and number of parameters, modality, license, technical means)
https://huggingface.co/CohereForAI/c4ai-command-r-v01 (includes intended tasks, use policy, release date, version, architecture and number of parameters, modality, license, technical means)
Example usage of the model cards: https://huggingface.co/HuggingFaceM4/idefics2-8b (includes intended tasks, use policy, release date, version, architecture and number of parameters, modality, license, technical means, dataset for training (see below as IDEFICS))
Documentation of SmolLM, including experiments, evaluation, and description of the training of the model: https://huggingface.co/blog/smollm
OpenLLM leaderboard, evaluation on standard evaluation metrics:
https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard
Existing events for red teaming:
https://aivillage.org/generative%20red%20team/generative-red-team-2/
The Environmental Impacts of AI – Primer:
https://huggingface.co/blog/sasha/ai-environment-primer
Energy Star Ratings for AI Models:
https://huggingface.co/blog/sasha/energy-star-ai-proposal

4. Beyond the minimal set of elements listed in the previous question, are there **other 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

Links to relevant material

Social impact evaluations, see Evaluating the Social Impact of Generative AI Systems in Systems and Society https://arxiv.org/abs/2306.05949

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

Measures to comply with rights reservation need to be accessible to SMEs and developers of open models, including actors leveraging public datasets. This requires ensuring that

opt-outs are publicly accessible, machine-readable with a known protocol. Such an approach also benefits rights holders more than fragmented protocols from GPAI developers. Measures to prevent the generation of copyright infringing content should also be cognizant of open developers and of the current lack of accepted definition of “substantial similarity” for model generations. A sufficiently detailed training data summary template is more accessible to open developers than under-defined output filtering requirements.

Links to relevant material

Open Future: Considerations For Implementing Rightholder Opt-Outs By AI Model Developers https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf

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
- ☐ I don't know

Your comments

700 character(s) maximum

The policy should ensure that the technologies and standards used for expressing opt-outs are widely and freely accessible, allowing rights holders and AI developers to engage without barriers that could adversely impact competition across developers. The opt-out process must be straightforward and user-friendly, avoiding the need for legal expertise and supporting automated processing to streamline compliance with the text and data mining rights reservation under Article 4(3) of Directive (EU) 2019/790. Additionally, tools to aggregate and manage opt-out requests are needed, as current solutions are insufficient and fail to prevent data from being crawled despite opt-outs.

Links to relevant material

Domain-level
Am I in The Stack? App https://huggingface.co/spaces/bigcode/in-the-stack
Hugging Face has integrated the Spawning HaveIBeenTrained API with the Hub for datasets that have an image_url field, e.g. see the report widget on the right in https://huggingface.co/datasets/kakaobrain/coyo-700m
CommonCrawl, which is a dataset of web crawl data, often used as the base for training data for models, respects opt-outs via robot.txt: https://commoncrawl.org/ccbot
robots.txt; ai.txt https://spawning.ai/ai-txt
TDM Reservation protocol (TDMRep), https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240202/

Work-level
Coalition for Content Provenance and Authenticity (C2PA)), https://c2pa.org/
International Standard Content Code (ISCC), http://iscc.codes/
Spawning.ai's Have I been trained?, haveibeentrained.com

Other relevant material
Analysis of opt-outs: https://www.dataprovenance.org/Consent_in_Crisis.pdf

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.

- ☑ Public/ open data repositories
- ☑ Content/data publicly available online (e.g. scraped from the internet)
- ☑ Proprietary data generated by the provider
- ☑ User-generated data obtained through the services or products provided by the provider
- ☑ Copyright protected content licensed by rightsholders

- ☑ 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.)
- ☑ 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

Information would ideally be detailed enough for any rights holder to understand how their data contributes to the GPAI system. Information needs to be detailed enough to ensure that broad trends in the use of personal data, creative works, and representations of specific groups affect the system are legible to external stakeholders including journalists and policymakers. This level of granularity requires different definitions for different categories, e.g. a list of URLs for public data repositories, the top domains by content in a web crawl, or the respective sizes and status of intermediaries (e.g. publisher, platform, archive) for different modalities of licensed content.

Links to relevant material

Open Future's Towards robust training data transparency:
https://openfuture.eu/publication/towards-robust-training-data-transparency/
What's in my big data: https://arxiv.org/abs/2310.20707 https://wimbd.apps.allenai.org/

On training data memorisation and PII data leakage:
Lukas et al.: Analyzing Leakage of Personally Identifiable Information in Language Models
https://arxiv.org/pdf/2302.00539
Carlini et al.: Extracting Training Data from Large Language Models
https://arxiv.org/abs/2012.07805

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.

- ☑ Modalities / type of data (text, images, videos, music, etc);
- ☑ Nature of the data (personal, non-personal or mixed);
- ☑ Time of acquisition/collection of the data;
- ☑ Data range of the data (e.g. time span), including date cutoffs
- ☑ In case of data scraped from the internet, information about the crawlers used;
- ☑ Information about diversity of the data (for example linguistic, geographical, demographic diversity);
- ☑ Percentage of each of the main data sources to the overall training/fine-tuning;
- ☑ Legal basis for the processing under Union copyright law and data protection law, as applicable;
- ☑ 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

All categories outlined have direct bearing on EU regulations including intellectual property, anti-discrimination, and personal data protections. While they can easily be documented for newly curated datasets, retroactive information gathering is more challenging. Requirements should reflect this by permitting less detailed documentation on pre-existing datasets, as long as they are properly identified and updates are fully documented. Requirements should be more limited for open access datasets and SMEs to reflect their inherent transparency and different resources respectively. The legal basis dimension can be the most difficult to assess and would benefit from harmonised interpretations.

Links to relevant material

Open Future's Towards robust training data transparency:
https://openfuture.eu/publication/towards-robust-training-data-transparency/

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

The performance of a GPAI system depends not only on the source of the data but on the specific learning objectives, data processing, and increasingly on techniques of “self-play” and RLHF whose specifics remain beyond the scope of the proposed summary. The latest release of the OpenAI o1 system in particular exemplifies the role of these advanced techniques in gaining a competitive advantage. Therefore, focusing information requirements on data that has external rights holders can help support the rights of EU citizens and support a healthy and sustainable data ecosystem without requiring GPAI developers to reveal their unique contributions to the performance of their products.

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.
☐

The list is comprehensive, as long as the risks currently listed under other systemic risks are prioritised alongside the other two categories; in particular safety, democratic processes, public and economic security, fundamental rights have been at the forefront of recent discussions such as the use of AI to decide whether unemployed workers get benefits [1] or access to information about democratic processes [2].

[1] https://gizmodo.com/googles-ai-will-help-decide-whether-unemployed-workers-get-benefits-2000496215
[2] https://www.latimes.com/opinion/story/2024-03-08/primaries-voting-elections-ai-misinformation-plaforms-chatgpt

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.
- ☐ 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.
- ☑ Other

Please specify

700 character(s) maximum

The content of the training datasets is a strong factor for most of the systemic risks considered. The prevalence of personal data, hate speech, information pertaining to capabilities of concern for the model, or known misinformation in the training dataset, for example, have direct bearing on the model behaviour; often more so than the results of a model on general performance benchmarks.

Your comments

700 character(s) maximum

Distribution strategies are important but equating open-source with higher risk is misguided. There is no evidence suggesting that open-source models are riskier than closed-source ones

[https://crfm.stanford.edu/open-fms/]. In fact, closed models are often more accessible due to user-friendly interfaces and cheap commercial access for a large user base [https://tinyurl.com/yf2yu8zv]; which compounds with the frailty of deployment-level guardrails [https://llm-attacks.org/].
Open-source releases offer unique benefits and risk strategies especially at the ecosystem level [https://huggingface.co/blog/ethics-soc-3]
Limitations of computation thresholds should be acknowledged [https://tinyurl.com/mrb4sjzc]

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”.

Question12. 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

Risk assessment measures should be developed in collaboration between actors of all sizes and allowing participation of affected stakeholders outside of the largest developers. This both ensures that the more meaningful risks are prioritised and allows smaller companies to follow established and scientifically validated standards, rather than having each actor reproduce work or prioritise risks at their own discretion without input from the most affected parties. This ensures that SMEs and start-ups can fully participate in responsible development.

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		X			
Forecasting model capabilities and risks before and during model development			X		
Continuous monitoring for emergence of risks, including data from users, relevant stakeholders, incident databases or similar			X		
Determining effectiveness of risk mitigation measures		X			
Safety cases to demonstrate that the model does not exceed maximum risk thresholds		X			
Aggregate risk assessment before model development		X			
Aggregate risk assessment before model deployment		X			
Aggregate risk assessment along the entire model lifecycle		X			

Third-party involvement in risk assessment, for example, related to inspections of training data, models or internal governance		X			

Your comments
700 character(s) maximum

Third-party involvement in risk assessment is essential, particularly for inspections of training data, models, or internal governance. In order for this feedback to be meaningful, the entire development cycle of GPAIs should be sufficiently meaningfully documented, allowing external stakeholders to direct their attention to the most relevant development choices. This transparency may be supported through documents like governance cards [https://arxiv.org/abs/2312.03872] to allow continuous external input.

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.

Evaluating the Social Impact of Generative AI Systems in Systems and Society https://arxiv.org/abs/2306.05949
Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law https://researchrepository.wvu.edu/wvlr/vol123/iss3/4/
The BigCode Project Governance Card: https://arxiv.org/abs/2312.03872

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**	X				
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		X			
Ensuring evaluations inform model deployment in real-world conditions	X				
Ensuring evaluations provide insights into the degree to which a model introduces or exacerbates risks			X		
Using non-public model evaluations, as appropriate		X			
Involve independent external evaluators, including with appropriate levels of access to the model and related information		X			
Involve affected persons, to understand effects of human interactions with a particular model over time		X			
Documenting evaluation strategies and results	X				

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

Your comments
700 character(s) maximum

Documentation/Reporting are core to ensuring accountability, the appropriateness of evaluations, and the development of new evaluation strategies.

Organisations who release open models, especially academic and SMEs, have different constraints and might not be able to hire external evaluators or report results privately. However, as models are de facto accessible to third parties, open source models as defined in the AI Act should be assumed to comply with these requirements.

Non-public evaluations should only be used to avoid data contamination, as transparency is necessary to build trust and accountability. Involving affected persons should be done with care.

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

On the importance of reporting evaluation results in detail:
Burnell et al.: Rethink reporting of evaluation results in AI; Aggregate metrics and lack of access to results limit understanding https://www.science.org/doi/10.1126/science.adf6369
Summary: https://montrealethics.ai/rethink-reporting-of-evaluation-results-in-ai/

On data contamination in evaluation data:
Balloccu et al.: Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in

Closed-Source LLMs https://arxiv.org/abs/2402.03927
Jacovi et al.: Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks https://aclanthology.org/2023.emnlp-main.308.pdf

On involving affected persons:
Sloane et al.: Participation Is not a Design Fix for Machine Learning
https://dl.acm.org/doi/pdf/10.1145/3551624.3555285
Birhane et al.: Power to the People? Opportunities and Challenges for Participatory AI
https://arxiv.org/pdf/2209.07572

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

While general-purpose AI model providers have access to significant technical knowledge and computational resources, they typically lack the expertise necessary to address social and economic risks of deploying models at scale – especially for risks that depend not just on properties of the model, but on its deployment and commercialisation formats.

Engaging external stakeholders and affected communities is essential for identifying many of the most current and acute risks of the technology, but care must be taken to ensure that their involvement is not exploitative and that they benefit from the process rather than being used solely for the improvement of AI models and systems.

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”.

Question18. 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

SMEs, start-ups, and open-source developers can implement technical risk mitigation by adhering to established good practices in AI model development. These interventions can be both more robust and more accessible than post-hoc interventions based on, e.g., safety fine-tuning or input filtering. Prioritising safety by design [1] and emphasising transparency

requirements help ensure that risk management is more accessible and achievable for providers of all sizes.

[1] https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.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	X				
Model design and development to achieve an appropriate level of trustworthiness characteristics such as model reliability, fairness or security			X		
Fine-tuning for trustworthiness and alignment such as through Reinforcement Learning from Human Feedback (RLHF) or Constitutional AI			X		
Unlearning techniques such as to remove specific harmful capabilities from models				X	

Technical deployment guardrails, such as content and other filters, capability restrictions, fine-tuning restrictions or monitoring-based restrictions in case of misuse by users			X		
Mitigation measures relating to the model architecture, components, access to tools or model autonomy	X				
Detection, labelling and other measures related to AI-generated or manipulated content		X			
Regular model updates, including security updates			X		
Measuring model performance on an ongoing basis			X		
Identification and mitigation of model misuse		X			
Access control to tools and levels of model autonomy			X		

Your comments
700 character(s) maximum

Reinforcement fine-tuning reduces unwanted model responses, but is brittle towards intentional misuse. Technical deployment guardrails, while helpful, have inherent limitations and must be thoughtfully implemented, taking into account deployment context. For open models, guardrails are often controlled by the deployer, not the developer. Dataset-level measures, like data governance, consistently mitigate risks and apply across both API deployment and open source. We also broadly caution against interventions that are beyond the state of the art and are not scientifically validated, such as unlearning or requiring fully reliable watermarking across modalities.

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.

Preventing misuse of models when developing them:
Kaffee et al.: Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing https://aclanthology.org/2023.findings-emnlp.932/

Jailbreaking of closed models:
Zou et al.: Universal and Transferable Adversarial Attacks on Aligned Language Models https://llm-attacks.org/ and https://arxiv.org/abs/2307.15043
Lu et al.: Set-level Guidance Attack: Boosting Adversarial Transferability of Vision-Language Pre-training Models https://openaccess.thecvf.com/content/ICCV2023/papers/Lu_Set-level_Guidance_Attack_Boosting_Adversarial_Transferability_of_Vision-Language_Pre-training_Models_ICCV_2023_paper.pdf

21. What are the greatest challenges that a general-purpose AI provider could face in implementing technical risk mitigation measures?
700 character(s) maximum

When deploying novel GPAI models, it is impossible to predict all possible misuses of the models. Therefore, adapting models for specific settings makes it easier to develop context-specific risk mitigation techniques [https://ieeexplore.ieee.org/abstract/document/1507050]. With wider access to the models through open models, domain experts are able to test for specific risks. It also requires updating of the risk mitigation strategies as they're developed.

Requiring open model providers to track all model reuse is technically infeasible as well as undesirable, as it restricts reuse possibilities for startups and research initiatives, limiting innovation and further development.

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

Clear open and accessible standards that can be followed without involving external organisations are much more accessible to SMEs and start-ups. Risk evaluation should avoid over-relying on mandatory external audits besides the one conducted through the AI Office or similar entities, and instead focus on clear, straightforward guidelines that small teams or individual developers can easily follow. The Code of Practice should not impose requirements that necessitate dedicated personnel, recognising the constraints faced by smaller entities.

Links to relevant material

Marino et al.: Compliance Cards: Automated EU AI Act Compliance Analyses amidst a Complex AI Supply Chain https://arxiv.org/abs/2406.14758

On the limitations of AI audits
Birhane et al.: AI auditing: The Broken Bus on the Road to AI Accountability https://arxiv.org/abs/2401.14462
Raji et al.: Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance https://dl.acm.org/doi/pdf/10.1145/3514094.3534181
Costanza-Chock et al.: Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem https://arxiv.org/pdf/2310.02521

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		X			
Internal independent oversight functions in a transparent governance structure, such as related to risks and ethics			X		

Traceability in relation to datasets, processes, and decisions made during model development	X				
Ensuring that staff are familiar with their duties and the organisation's risk management practices		X			
Responsible scaling policies		X			
Acceptable use policies	X				
Whistleblower protections	X				
Internal resource allocation towards risk assessment and mitigation measures as well as research to mitigate systemic risks		X			
Robust security controls including physical security, cyber security and information security		X			
External accountability measures such as third-party audits, model or aggregated data access for researchers	X				
Other collaborations and involvements of a diverse set of stakeholders, including impacted communities		X			
Responsible release practices including staged release, particularly before open-sourcing a model with systemic risk		X			

Transparency reports such as model cards, system cards or data sheets	X				
Human oversight mechanisms		X			
Know-your-customer practices		X			
Logging, reporting and follow-up of near-misses along the lifecycle			X		
Measures to mitigate and remediate deployment issues and vulnerabilities		X			
Complaints handling and redress mechanisms, such as bug bounty programs	X				
Mandatory model updating policies and limit on maximum model availability			X		
Third-party and user discovery mechanisms and reporting related to deployment issues and vulnerabilities		X			

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.

Complaints handling and redress mechanisms:
Cattell et al.: Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities
https://arxiv.org/abs/2402.07039

Staged releases:
Solaiman: The Gradient of Generative AI Release: Methods and Considerations
https://arxiv.org/abs/2302.04844

On involving affected communities:
Sloane et al.: Participation Is not a Design Fix for Machine Learning

https://dl.acm.org/doi/pdf/10.1145/3551624.3555285
Birhane et al.: Power to the People? Opportunities and Challenges for Participatory AI
https://arxiv.org/pdf/2209.07572

25. What are the greatest challenges that a general-purpose AI provider could face in implementing governance risk mitigation measures?
700 character(s) maximum

Smaller providers, such as SMEs and startups, often lack the financial and human resources needed to establish dedicated governance structures and devise their own risk management frameworks from scratch. The rapidly changing nature of AI and its applications requires governance measures to continuously adapt to new risks and developments. Meaningfully involving diverse stakeholders, including impacted communities, in governance practices requires effective communication and collaboration. Ease of implementation by smaller actors and of adaptation to changing conditions for all can be greatly facilitated by direct collaboration between actors.

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.

Question26. 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

The Code of Practice aims to ensure that developers demonstrate adequate foresight and

consideration. This can be validated in collaboration with external bodies, such as the scientific panel. When alerts arise regarding the use of GPAI models, the panel should evaluate the extent to which the Code of Practice was followed and whether it adequately covered and addressed the issues.
The Code of Practice is considered successful if it meets the dual goals of seeing extensive and transparent adoption from all categories of developers and of serving as an effective support for assessing practices in the light of stakeholder complaints to support more reliable and trustworthy technology.

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

Key performance indicators and reporting commitments for providers should reflect differences in size by tailoring complexity and resource requirements to the provider's scale. SMEs and startups could have simpler reporting processes and more flexible timelines, with access to open-source tools to reduce costs. For example, smaller entities may use internal evaluations or peer reviews instead of costly third-party audits. KPIs should focus on core issues like data governance, allowing startups to gradually adopt standards while remaining competitive, fostering innovation, and encouraging diverse ecosystem participation.

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".

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.	Rank 2
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	Rank 1

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	Rank 3

Your comments

700 character(s) maximum

The AI Office should incorporate expert feedback on the Code of Practice and make necessary adjustments based on this input. Nevertheless, regular reviews are essential to determine if updates are needed. While foresight can be a useful tool to analyse current trends, in the context of the fast developing technology of AI it is not a proven tool to implement policy changes.

Link to relevant material

A way of enabling the general public to alert, e.g., the AI Office of model flaws and risks are coordinated flaw disclosures: Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities https://arxiv.org/abs/2402.07039