General-Purpose AI Code of Practice: **AI Standards Lab** Survey feedback

Fields marked with * are mandatory.

About this document:

This is the survey response text of **AI Standards Lab,** for the first draft of the EU GPAI Codes of Practice, which was sent out to participants by the EU AI Office. We are publishing our responses to support transparency and cooperation.

For any questions or comments, please email inquiries@aistandardslab.org

Some sections were omitted due to the CoP privacy code of conduct - the remaining sections include only information explicitly mentioned in the public first draft. Blank response fields were also removed.

For a public copy of our Codes of Practice Consultation submission (referenced in some of the text boxes), see our website.

Provide your feedback to the first General-Purpose AI Code of Practice!

Thank you for participating in the drawing-up of the first General-Purpose AI Code of Practice.

Upon receiving the first draft, you are encouraged to express your comments on the content via this survey, **deadline Thursday 28 November 2024, 12:00 CET**.

Your feedback is essential in helping us understand how the Code of Practice can best serve and support stakeholders across diverse sectors, leading to a final Code of Practice which should reflect the different submissions as far as possible, while ensuring a convincing implementation of the legal framework. Please be aware that the survey does not cover Art. 53(1)(d) issues.

For each section/measure/sub-measure of the Code of Practice, participants will be asked to answer **two types of questions**:

1. **Opinion rating** (close-ended feedback): express the level of agreement with the content choosing among different options.
2. **Open-ended questions**: specific to each sub-section's measures and sub-measures, and additional questions cross-measures. This includes the opportunity to comment on each section and the overall draft.

In addition, you may upload supporting documents at the end of the survey.

Stakeholder information

Please provide your name, surname, email address, and the name of your organisation (if applicable).

Please note that if your contact information does not correspond to an eligible participant or to the

organisation's Point of Contact, your response will be discarded.

	Organisation (if applicable)
Stakeholder *	**AI Standards Lab**

Which stakeholder category would you (or your organisation) consider yourself in? *

- ○ Academia (in a personal capacity)
- ○ Civil society organisation
- ○ Downstream provider of an AI system based on general-purpose AI models, or acting on behalf of such providers
- ○ EU Member State representative
- ○ European or international observer
- ○ Other independent expert (in a personal capacity)
- ○ Other industry organisation, or acting on behalf of such organisations
- x ○ Other organisation with relevant expertise
- ○ Other stakeholder organisation
- ○ Provider of a general-purpose AI model, or acting on behalf of such providers
- ○ Rightsholder organisation

Please indicate all the working groups you participate in. Please note that if you are the Point of Contact of your organisation, you should select all the working groups of your representatives. *

- x ☐ Working Group 1: Transparency & copyright-related rules
- x ☐ Working Group 2: Risk identification and assessment for systemic risk
- x ☐ Working Group 3: Technical risk mitigation for systemic risk
- x ☐ Working Group 4: Governance risk mitigation for systemic risk

Please indicate which section you wish to provide your feedback. If you wish to comment on all sections, please select all the options. *

- x ☐ Overall Code of Practice Draft
- x ☐ Section II: [Working Group 1] Rules for providers of general-purpose AI models
- x ☐ Section III: [Working Group 2] Taxonomy of systemic risks
- x ☐ Section IV: [Working Groups 2/3/4] Rules for providers of GPAI models with systemic risk

Overall Code of Practice Draft

To what extent are you satisfied with the overall content of the Code of Practice?

- [ ] 1: Dissatisfied
- [ ] 2: Slightly dissatisfied
- [x] 3: Moderately satisfied
- [ ] 4: Mostly satisfied
- [ ] 5: Highly satisfied

Would you like to share any comments on the overall Code of Practice Draft? Please note that any feedback to specific content should be given per Section.

2000 character(s) maximum

The draft overall seems like a good start in terms of mirroring the GPAI provider obligations in the AI Act, but to become useful more details and KPIs need to be added. To add more detail efficiently, the chairs will need to set up a feedback process where they avoid introducing information flow bottlenecks and actively leverage the inputs and volunteer capacity of CoP participants. To do this, the first steps could be to share earlier consultation inputs, and inputs from this survey, and then invite participants to directly engage with each other to prepare written inputs for targeted sessions.

Section I (preamble) start of II (whereas) look pretty good, but next time please also provide an opportunity to give feedback on them.

We recommend:
- Add a section on Codes review, monitoring and enforcement that specifies review processes, monitoring requirements, and enforcement powers.
- Define and explain terms, take care to not re-define terms defined by the AI Act but copy the AI Act definitions. See also our attached input on 'risk terminology' for more suggestions relating to the definition and disambiguation of 'risk' as it appears in the AI Act.
- Add a lot more detail to the risk taxonomy based on consultation inputs. Our own earlier consultation input includes short descriptions of many of the systemic risk sources and various risk management methods, intended for direct cut-and-pasting.
- Clarify qualifiers such as "reasonable," "meaningful,". See the attached input 'Feedback on clarifying measures with algorithms' for more detailed insights on how to do this.
- Clarify the CoP implementation process for a GPAI model provider - which steps should be done during the model development/deployment, or are model-agnostic (organization level)l. See attached document "Feedback on Implementation of the Code."
- The SSF should detail both proactive and reactive mitigation measures. Currently, reactive measures are only briefly mentioned under Measure 18.

Section II: [Working Group 1] Rules for Providers of General-Purpose AI Models

Measures/Sub-measures Specific Feedback on Section II: [Working Group

1] Transparency

In this section you are asked to provide your overall opinion on the measures and sub-measures included in the second section of the Code of Practice related to *Transparency*.

Transparency, Measure 1: Documentation for the AI Office

To what extent do you agree with this measure?

- ☐ The measure should be removed in its entirety
- ☐ The measure should be substantially edited and/or further clarified
- X ☐ The measure should be lightly edited and/or further clarified
- ☐ The measure is close to where it needs to be

Please explain your rating to this measure

2000 character(s) maximum

More details on evaluation should be provided and clarified.

What KPI would you add for this measure?

2000 character(s) maximum

For Article 53 (1) (a), add the metrics from stated in the initial consultation: "*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"*

And "Evaluation strategies shall include evaluation criteria, metrics and the methodology on the identification of limitations"

On top of describing the methodology on the identification of limitations, GPAI providers should report on their ongoing steps in addressing these limitations, such as their current efforts on developing new evaluations for their GPAIs, including (1) on what areas will they release concrete evaluations soon and (2) on what areas they hope or struggle to develop evaluations.

Transparency, Measure 2: Documentation for downstream providers

To what extent do you agree with this measure?

- ☐ The measure should be removed in its entirety
- ☐ The measure should be substantially edited and/or further clarified
- X ☐ The measure should be lightly edited and/or further clarified
- ☐ The measure is close to where it needs to be

Please explain your rating to this measure

2000 character(s) maximum

For Annex XI §1 2.(c) and Annex XII 2.(c), the CoP states:
Information on data used for training, testing and validation: "Signatories should detail the data acquisition methods, specific information for each data acquisition method"

- This information is set to be available to both AIO and downstream providers. However, the current phrasing is unclear on what “specific information” means.

Please provide suggestions on how to improve this measure

2000 character(s) maximum

For Annex XI §1 2.(c) and Annex XII 2.(c), the CoP states:
“Signatories should further detail <...> the methods used to detect unsuitability of data sources and any biases in the data.”
-Currently, it is set to be disclosed only to the AIO. This should be disclosed with the downstream providers too to support downstream providers' debiasing efforts of their version of the model.

For Annex XI §1 1.(c) and Annex XII 1.(c), the code should document “Mandatory model updating policies and limit on model availability” as specified by the provider.

For the items listed in the table (at page 10), how should the Code of Practice provide greater detail? *2000 character(s) maximum*

For *Information on data used for training, testing and validation* used in *Annex XI §1 2.(c) and Annex XII 2.(c):*

A.Data sources, data acquisition and other methods of generating data - Include the type of source, but not limited to:

1.Web-scraped data: domain information for the first 100K most-scraped domains or Top 5%, whichever is higher
2.Licensed works: include the source and sub-type (exclusively licensed or not)
3.Synthetic data & provider-generated data: include generation methods

B.Specific use of each dataset in the model development pipeline (training, testing, validation, fine-tuning, reward modeling, etc.)

Further details can be found in the Annex 1 of: Warso, Z., Gahntz, M., & Keller, P. (2024). *Sufficiently detailed? A proposal for implementing the AI Act’s training data transparency requirement for GPAI*.
https://openfuture.eu/wp-content/uploads/2024/06/240618AIAtransparency_template_requirements-2.pdf

The Documentation as a whole should be presented in a readable and widely accessible format incorporating the state of the art, such as model cards.

For which of the topics below is more clarification or specificity most needed? (*select all that apply*)

- [ ] General information
  - [ ] Intended uses
  - [ ] Acceptable use policies
  - [X] Methods of distribution
  - [ ] Interaction with hardware and software
  - [ ] Software versions
  - [X] Model architecture and parameters
  - [ ] Input and output modalities
  - [X] License
  - [X] Technical means for downstream integration

- X ☐ Training process
- X ☐ Training, testing, validation data
- X ☐ Computational resources
- X ☐ Energy consumption
- X ☐ Testing process

If you are *not* a General-Purpose AI Model provider, for which of the topics below would you prefer information be encouraged to be made public? (*select all that apply*)

- ☐ Not applicable
- X ☐ General information
- X ☐ Intended uses
- X ☐ Acceptable use policies
- ☐ Methods of distribution
- X ☐ Interaction with hardware and software
- X ☐ Software versions
- ☐ Model architecture and parameters
- X ☐ Input and output modalities
- X ☐ License
- X ☐ Technical means for downstream integration
- X ☐ Training process
- X ☐ Training, testing, validation data
- ☐ Computational resources
- ☐ Energy consumption
- X ☐ Testing process

WG1 - Section II - Measures/Sub-measures Specific Feedback on: [Working Group 1] Copyright-related rules

In this section you are asked to provide your overall opinion on the measures and sub-measures included in the second section of the Code of Practice related to *Copyright-related rules.*

Measure 3: Put in place copyright policy

To what extent do you agree with this measure?

- ☐ The measure should be removed in its entirety
- ☐ The measure should be substantially edited and/or further clarified
- ☐ The measure should be lightly edited and/or further clarified
- X ☐ The measure is close to where it needs to be

Sub-Measure 3.1: Draw up and implement a copyright policy

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [X] The sub-measure is close to where it needs to be

Sub-Measure 3.2: Upstream copyright compliance

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [X] The sub-measure is close to where it needs to be

Sub-Measure 3.3: Downstream copyright compliance

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [X] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

2000 character(s) maximum

The measure's language can be further simplified.

Please provide suggestions on how to improve this sub-measure

2000 character(s) maximum

Revise the penultimate sentence to: "Signatories are encouraged to make their model outputs avoid reproducing copyrighted work. In the case of contracts with downstream providers, signatories should compel downstream providers to avoid reproducing copyrighted work with the provider's model/s as a fundamental condition for the contract validity.

Measure 4: Compliance with the limits of the TDM exception

To what extent do you agree with this measure?

- [ ] The measure should be removed in its entirety
- [X] The measure should be substantially edited and/or further clarified
- [ ] The measure should be lightly edited and/or further clarified
- [ ] The measure is close to where it needs to be

Please explain your rating to this measure

2000 character(s) maximum

The rights to mine data, such that the data can be processed for certain purposes, are also limited by the GDPR, in case that the data contains personal/privacy sensitive information. It would be good if the Code includes information on measures that signatories/providers should take in light of GDPR compliance for themselves, as data controllers and/or processors, and further throughout the value chain. In particular, measures 4.3 and 4.4 could be expanded (or separate measures could be added) on the topic of representing and respecting statements of consent as are required to be sought under the GDPR.

Please provide suggestions on how to improve this measure

2000 character(s) maximum

We recommend that the (vice) chairs explicitly seek input on this GDPR matter from CoP participants, specifically requesting input of text that might be cut-and-pasted directly into the CoP. If no such input is forthcoming, then (because of the short timing) we do not believe the (vice)chairs should try to write text themselves. If one or more inputs are received, they should be shared by the Chairs for comments.

Sub-Measure 4.1: Respect Robots.txt

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [X] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

2000 character(s) maximum

It does not explicitly handle the case of signatories buying crawled data and whether that crawled data needs to respect robots.txt. This creates an undesirable ambiguity in how the code should be interpreted.

Please provide suggestions on how to improve this sub-measure

2000 character(s) maximum

Explicitly declare that signatories that buy crawled data must ensure that the crawled data respects robots.txt in accordance with Sub-Measure 3.2.

Sub-Measure 4.2: No effect on findability

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [X] The sub-measure is close to where it needs to be

Sub-Measure 4.3: Best efforts regarding other appropriate means

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [x] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

2000 character(s) maximum

We believe that the language 'and comply with other appropriate machine-readable means to express a rights reservation' is an essential part of this measure, because unfortunately processes to create widely accepted industry standards via a consensus process in this field have often failed.

Please provide suggestions on how to improve this sub-measure

2000 character(s) maximum

Change wording 'industry standards' to say 'industry standards and de-facto standards'.

Invite CoP participants to submit examples of existing 'other appropriate machine-readable means' widely used now, add these to the code. Some participants might have already submitted such means.

Add a clarifying statement as follows 'Web pages, and or documents shared on the web like PDF documents, often contain in human-readable text expressing a rights reservation, e.g. a reference to a standardised license conditions like a creative commons license. This text is often formatted in a way so that it can be easily located and decoded by text analysis software. For the purpose of this measure, such statements in text that can be easily located and decoded by software are to be considered machine-readable means to express a rights reservation.

Sub-Measure 4.4: Commitment to collaborative development of rights reservations' standards

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [x] The sub-measure is close to where it needs to be

Sub-Measure 4.5: No crawling of piracy websites

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [x] The sub-measure is close to where it needs to be

Measure 5: Transparency

To what extent do you agree with this measure?

- [ ] The measure should be removed in its entirety
- [ ] The measure should be substantially edited and/or further clarified
- [ ] The measure should be lightly edited and/or further clarified
- [X] The measure is close to where it needs to be

Sub-Measure 5.1: Public information about rights reservation compliance

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [X] The sub-measure is close to where it needs to be

Sub-Measure 5.2: Crawler name and robots.txt features

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [X] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

2000 character(s) maximum

Merely providing the crawler name is insufficient to determine whether rights of rightsholders were respected. Rightsholders may not have sufficient technical expertise to understand what domains were scrapped by each crawler, let alone which crawlers are proprietary and whose further details are unknown.

Furthermore, embracing robots.txt as the central data protection mechanism for transparency ignores the possibility of alternative mechanisms being developed and used.

Obligation to disclose crawler names should also apply to data sets that the signatory has bought from third parties.

Please provide suggestions on how to improve this sub-measure

2000 character(s) maximum

Add the possibility of using different data protection mechanism aside from/ in addition to robot.txt, if decided by the AI Office at a later time.

Add text that Obligation to disclose crawler names should also apply to a) data sets signatory obtains from third parties, or b) that have been used to create a model that signatory obtained from third party, where that model is b1) either the basis of signatories' own model, or b2) used to tune or train the signatories' own model.

What KPI would you add for this sub-measure?

2000 character(s) maximum

For each crawler used, list the first 100K most-scrapped domains or the top 5% domains, whichever is larger.

Sub-Measure 5.3: Single point of contact and complaint handling

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- X ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

2000 character(s) maximum

Complaint handling mechanism needs to be specified further and must uphold the practical interests of both the rightholders and the providers.

Please provide suggestions on how to improve this sub-measure

2000 character(s) maximum

Include a provision of a target case resolution to prevent a backlog of issues submitted by rightsholders and thus paralyzing the complaint handling mechanism.

What KPI would you add for this sub-measure?

2000 character(s) maximum

Set a target of 3 months or less for resolution of 85% or more issues requested.

Sub-Measure 5.4: Documentation of data sources and authorisations

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- X ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

2000 character(s) maximum

This sub-measure is related to Measure 1 but no reference is made to it

The sub-measure does not adequately mirror article 53(1)(d) which requires certain data to be made publicly available.

Please provide suggestions on how to improve this sub-measure

2000 character(s) maximum

Harmonize this measure with the requirements set out in Measure 1, and possibly add further on top of it. Namely, clarify the specific steps providers need to make to disclose the use of protected content used in their model development. At the very least, this measure should refer to Measure 1.

Add language (or a new measure) mirroring article 53(1)(d). Alternatively, update point b) in preamble to explicitly state that the Code does not give any guidance on article 53(1)(d). Our concern is here that the Code, if confirmed by an implementing act, would otherwise weaken or make ambiguous recital 107 of the Act – such ambiguity would in our analysis have a negative impact on copyright holders and the functioning of the GPAI market.

Section III: [WG2] Taxonomy of Systemic Risks

Taxonomy of systemic risks, Measure 6: Taxonomy

To what extent do you agree with this measure?

- ☐ The measure should be removed in its entirety
- x ☐ The measure should be substantially edited and/or further clarified
- ☐ The measure should be lightly edited and/or further clarified
- ☐ The measure is close to where it needs to be

Please provide suggestions on how to improve this measure

500 character(s) maximum

In our initial consultation input, we proposed that the taxonomy should include an extensive catalog of risk sources, based at least on the material we included in our free-text submission. This same material is also presented in our public-domain paper on risk sources (Gipiškis et al., 2024). We volunteer to write additional text or adapt our submitted content as needed. Please also refer to our supporting document, "Feedback on risk terminology and risk sources."

What KPI would you add for this measure?

500 character(s) maximum

GPAI model providers may adopt the risk taxonomy from Measures 6.1, 6.2, and 6.3 in its entirety. Alternatively, they may develop their own set of risk taxonomy, which includes risks listed in Measures 6.1, 6.2, and 6.3 as a minimum, that appropriately captures the types, natures, and sources of systemic risks.

Sub-Measure 6.1: Types of systemic risks

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- x ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

a) As it defines 'types of systemic risk,' the text under 6.1 can easily be interpreted as overruling the types of 'systemic risk' listed in Article 3(65), where systemic risk is defined as the probabilities and severity of certain specific harms: this creates major legal/interpretation uncertainties. b) adding much more detail to the taxonomy would serve the goals of the code better.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

Disambiguate the taxonomy text in the Code so that it becomes very clear that it does not override or replace Article 3(65).

The types of systemic risks should reflect those listed in Article 3(65) and should be distinct from the nature and sources of systemic risks. We elaborate further in the attached document, "Feedback on Risk Taxonomy." See also our attached input on 'risk terminology' for more suggestions relating to the definition of 'risk' as it appears in the AI Act.

What are relevant considerations or criteria to take into account when defining whether a risk is a systemic risk?

500 character(s) maximum

We believe that all of the considerations and criteria are already included in Article 3(65) of the Act, namely (i) specific to the high-impact capabilities of general-purpose AI models, (ii) having a significant impact on the Union market due to their reach, and (iii) can be propagated at scale across the value chain.

Based on these considerations or criteria, which risks should be prioritised for addition to the main taxonomy of systemic risks?

500 character(s) maximum

We advise against prioritizing specific risks explicitly (and rather mainly having the list serve as guidance), as this is contrary to best practices in safety engineering. In GPAI , the risk landscape may change quickly, and may be directly informed by internal GPAI provider information, where they are most well-equipped to discover and prioritize risks. The responsibility of identifying and prioritizing risks should be on the GPAI provider, with external stakeholders serving as guidance.

Sub-Measure 6.2: Nature of systemic risks

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- x ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

Additional dimensions (e.g., affected entities, uncertainty levels, and interdependencies) as well as more corresponding examples might be added to provide more details.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

For 6.2 Nature of systemic risks, actors driving the risk could include multi-agent systems.

Sub-Measure 6.3: Sources of systemic risks

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- x ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

We propose adding a lot more content here to improve the usability of the Code. Individual risk sources are preferably not described with a single line, but with a title followed by one or more descriptive paragraphs. We would be happy to assist the (vice)chairs in adding content, based on our consultation submission or other sources, if invited to do so.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

We recommend consulting our initial free-text submission to the multi-stakeholder consultation for more detailed descriptions of various sources of systemic risks to be added to the taxonomy 6.3. See our attached input "Feedback on risk terminology and risk sources" for more details.

We further recommend adding items "superhuman speed of operation," "resistance to shutdown" and "lack of recallability (in open source models)."

Section IV: [Working Groups 2/3/4] Rules for Providers of General-Purpose AI Models with Systemic Risk

Section IV: [Working Group 2] Risk assessment for providers of General-Purpose AI Models with Systemic Risk

In this section you are asked to provide your overall opinion on the measures and sub-measures included in the fourth section of the Code of Practice related to *Risk assessment for providers of General-Purpose AI Models with Systemic Risk.*

Measure 8: Risk identification

To what extent do you agree with this measure?

- [ ] The measure should be removed in its entirety
- [ ] The measure should be substantially edited and/or further clarified
- [x] The measure should be lightly edited and/or further clarified
- [ ] The measure is close to where it needs to be

Please explain your rating to this measure

500 character(s) maximum

Additional details need to be added to the description.

Please provide suggestions on how to improve this measure

500 character(s) maximum

The description should be expanded to include not only the identification of systemic risks but also their sources (Art. 55(1)(b)).

What KPI would you add for this measure?

500 character(s) maximum

(1) Early Warning Effectiveness Rate (adapted from financial risk management)
(2) Percentage of Risks Identified Before Manifestation
(3) Is there real-time monitoring of model behavior?
(4) Are critical failures logged and tracked?
(5) Are monitoring tools regularly updated?

Sub-Measure 8.1: Determining risks

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [x] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

Additional details need to be added to the description.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

A) The description should be expanded to include not only the identification of systemic risks but also their sources (Art. 55(1)(b)).

B) Remove “particularly” from the first sentence in the description.

C) Change “they will use the systemic risks listed in the taxonomy” to “they will use the taxonomy.”

Measure 9: Risk analysis

To what extent do you agree with this measure?

- ☐ The measure should be removed in its entirety
- ☐ The measure should be substantially edited and/or further clarified
- x ☐ The measure should be lightly edited and/or further clarified
- ☐ The measure is close to where it needs to be

Please provide suggestions on how to improve this measure

500 character(s) maximum

(Vice)chairs could consider entirely deleting measures 9.2 and 9.3, as an alternative to attempting to clarify what 'risk indicators' and 'tiers of severity' really are. From the experience of one of us in JTC21, it is time-consuming to define specific process elements in a way that are general enough, and it is better to focus energy on defining high level outcomes required of methodologies (as in 9.1), not their detailed steps. In case they decide to not delete, see input below.

What KPI would you add for this measure?

500 character(s) maximum

(1) Identification of Risk Precursors (situations with components that could lead to undesirable outcomes)

Sub-Measure 9.1: Methodologies

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- x ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

More details could be added about the "robust risk analysis methodologies." (1) What precisely constitutes their "robustness"? (2) What criteria should be used to determine when a particular methodology should be used over the other? (3) The description should be expanded to also include the sources of the pathways to systemic risks.

What KPI would you add for this sub-measure?

500 character(s) maximum

(1) Were precursors to systemic risks identified?
(2) Were necessary and sufficient conditions for systemic risks identified?
(3) Do the probabilities assigned to pathways correspond to the frequency of risks materialising through those pathways?

Sub-Measure 9.2: Mapping to systemic risk indicators

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- x ☐ The sub-measure is close to where it needs to be

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

More explanation is needed on what is meant by "systemic risk indicators." A definition should be provided in the description.

Sub-Measure 9.3: Tiers of severity

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- X ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

The sub-measure description should specify whether providers would be defining these tiers themselves.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

The sub-measure should mention that in cases where the level of risk is intolerable and no appropriate safeguards are available, the model should not be deployed. The minimum requirement on any tier system used is that it can make this distinction when drawing conclusions.

Is 'severity' the best way to articulate levels of 'gravity' or could it create confusion with the definition of risk as the combination of probability and severity?

500 character(s) maximum

Severity should be kept. It should already include the scale of harm within it (for example, a single injury/death should be considered less severe than a large-scale event causing multiple casualties).

What will those tiers of severity be? Is there already a nascent standard or a consensus forming?

500 character(s) maximum

The Act implies that a binary distinction needs to be made between 'tolerable' and 'intolerable.' Given the diversity of types of harms and risk sources and analytical methods, no single standardised tier system could ever work to support analysis across this diversity. However, the code can recommend certain tier systems to be used for sub-cases in risk analysis.

Sub-Measure 9.4: Forecasting risks

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [x] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

More details need to be added. Vague wording needs to be improved or clarified.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

(1) More information on what "best effort estimates" entail. (2) Justifications for and reasoning behind the underlying assumptions in those estimates. (3) It might be worth considering adding confidence intervals. (4) Improved wording and an explanation of what it means to "trigger" an "indicator." (5) More details on what happens after the indicator has been triggered.

Measure 10: Evidence Collection

To what extent do you agree with this measure?

- [ ] The measure should be removed in its entirety
- [ ] The measure should be substantially edited and/or further clarified
- [x] The measure should be lightly edited and/or further clarified
- [ ] The measure is close to where it needs to be

What KPI would you add for this measure?

500 character(s) maximum

(1) Are areas with scarce or missing evidence being tracked? (2) Are evidence-collection methods documented?

Sub-Measure 10.1: Model-agnostic evidence

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [x] The sub-measure is close to where it needs to be

Sub-Measure 10.2: Best-in-class evaluations

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- X ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

The sub-measure states that "existing knowledge about the behavior of very similar models may, for example, reduce the depth of evaluation needed." However, one could imagine hypothetical scenarios where new capabilities emerge in one of the "similar" models but not in the other. The sub-measure should clarify the criteria that need to be used to evaluate this "similarity" and address the corresponding uncertainties.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

It is insufficient to let the provider choose whether evaluations can be done by internal or external evaluators, so clearer criteria are needed. External options should be independent third parties.

What factors might determine whether a certain evaluation method is an adequate fit for a specific model and risk, and whether an evaluation was thorough enough?

500 character(s) maximum

Currently, there are no SOTA solutions for this. Adequacy of evaluations can be assessed by how easily novel capabilities are found with further evaluation. Relevant factors for thorough evaluations include the time given to evaluators, the number of independent evaluators (which could pursue different lines of inquiry), the number of test cases per systemic risk, and the scores on well-known safety and alignment benchmarks, such as MACHIAVELLI and SafeBench.

Sub-Measure 10.3: Scientific rigour and other quality factor

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- X ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

Generally accepted SOTA risk assessment methodologies incorporate a broader range of reasoning than what is found in 'the scientific method only; therefore, saying 'high scientific rigour' can be read as giving signatories a free pass not to be rigorous in other assessment aspects, e.g. related to legal interpretations or fundamental rights impact assessments, or to discard mitigation measures which they judge as not scientifically rigorous, even if they are accepted risk management practices.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

Remove word 'scientific' in the phrase 'high scientific rigour', or use "safety engineering rigour." Add language clarifying what rigour means.

See the proposed measure called 'Document the risk assessment in a structured way', present in our consultation input, for language that explores and defines rigour in a more general and useful way. For convenience, we have reproduced this proposal, with further comments, in our attached input 'Feedback on rigour, scientific and other types.'

How should high scientific rigor be operationalised? What is the gold standard and when should Signatories deviate from it (for example when conducting early, exploratory research)?

500 character(s) maximum

See our attached input 'Feedback on rigour, scientific and other types' for more details.

Sub-Measure 10.4: Capability elicitation

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- X ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

The description could include an observation that capability elicitation may be dangerous and that signatories should ensure its safe implementation.

Sub-Measure 10.5: Models as part of systems

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- X ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

The boundary between model and system should be better defined, to bound/scope of GPAI provider responsibility. Some example criteria:
(1) direct enablement of risk by model provider (e.g., programming ability inherent to model leading to cyber misuse)

(2) effectiveness of system level mitigations (e.g., model generating harmful outputs or behaving deceptively is hard to mitigate sufficiently using only downstream provider output filtering)
(3) model level risks cascading to multiple systems

What KPI would you add for this sub-measure?

500 character(s) maximum

Have models been tested using best performing available AI systems?
Have providers of AI systems been consulted in this assessment (as appropriate - for example, providers of agentic coding assistants)?
Have boundaries of responsibility for model provider and downstream providers of AI systems been scoped?
Have risks of mis-scoping the boundary been identified?

Sub-Measure 10.6: Diverse evaluations & generalisation

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [X] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

More details need to be added to the description.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

Please refer to our free-text submission for Working Group 2 in the Consultation survey form, in particular heading "Model Evaluations" in WG2 and WG3. This input is also accessible in our public domain paper "Risk Sources and Risk Management Measures in Support of Standards for General-Purpose AI Systems" on arXiv for concrete inputs on evaluations, including diversity.

Sub-Measure 10.7: Exploratory work

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [X] The sub-measure is close to where it needs to be

Sub-Measure 10.8: Sharing tools & best practices

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety

- [ ] The sub-measure should be substantially edited and/or further clarified
- [x] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

More information is needed on "specifically identified cases" when "signatories may limit the sharing of information."

What channels, organisations and methods exist that would facilitate the sharing of evaluations, tools, and best practices, while not putting undue additional pressure on the research teams currently working at the cutting edge of AI Safety?

The UK AISI "Inspect" open source framework for language model evaluations. METR and Apollo Research are developing and sharing insights on evaluations for some of the capabilities/propensities in 6.3.1/2

Is this measure especially beneficial to startups and Small Medium Enterprises (SMEs) who might not have as much capacity to develop these tools and practices from scratch, but might be able to use them?

Yes, we think this can be very beneficial, especially as model training may take up a significant portion of an SME's budget, and may not have the capacity to develop their own frameworks.

Sub-Measure 10.9: Sharing results

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [x] The sub-measure is close to where it needs to be

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

The sub-measure could benefit from more details on what an "easily comparable format" would entail.

Measure 11: Risk assessment lifecycle

To what extent do you agree with this measure?

- [ ] The measure should be removed in its entirety
- [ ] The measure should be substantially edited and/or further clarified
- [x] The measure should be lightly edited and/or further clarified
- [ ] The measure is close to where it needs to be

Please explain your rating to this measure

500 character(s) maximum

We suggest light edits to sub-measures 11.2, 11.3, and 11.4, as detailed below.

Sub-Measure 11.1: Before training

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- X ☐ The sub-measure is close to where it needs to be

Sub-Measure 11.2: During training

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- X ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

"Regular milestones" should be more clearly defined in KPIs.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

We suggest that the four-fold increase be expanded by mentioning that very early in training the increase can be slower (since the first teraFLOP likely does not pose high risk), and later increases are more relevant (when the model acquires real-world capabilities). We agree that "training" should include changes to the weights of a model in order to increase its capabilities.

Sub-Measure 11.3: During deployment

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- X ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

The distinction between Sub-Measure 11.3 and 11.4 is not clear. If anything, Sub-Measure 11.3 should be about one-off processes, while 11.4 should be about continuous processes. However, Sub-Measure 11.3 also includes processes at the frequency of "at least every six months."

Sub-Measure 11.4: Post-deployment monitoring

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- X ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

The distinction between sub-measures 11.3 and 11.4 needs to be clarified.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

Suggest to add "Third-party and user discovery mechanisms and reporting related to deployment issues and vulnerabilities" as per the initial consultation survey question.

What methods exist (or could exist) that would enable providers of open-weights General-Purpose AI Models with Systemic Risk to monitor models they have released, without major side effects for the downstream users of these models?

Incident reporting by downstream users, such as cybersecurity norms of vulnerability reporting. Care should be taken to report to all affected model providers - many vulnerabilities are transferable - e.g. between all models trained on a given dataset, or with a similar architecture. A broader collaborative vulnerability reporting platform is preferable to company specific.

We are somewhat worried about what this question implies. Correct systemic risk mitigation might require measures that create substantial costs/effects for downstream model users. Adherence to these measures can be required in the open-weight model license. If an open-wright model provider believes there is a real possibility that their users will ignore the license in a way that creates intolerable systemic risk, they shall not release the model.

Section IV: [Working Group 3] Technical risk mitigation for providers of General-Purpose AI Models with Systemic Risk

In this section you are asked to provide your overall opinion on the measures and sub-measures included in the fourth section of the Code of Practice related to *technical risk mitigation for providers of General Purpose AI Models with Systemic Risk.*

Measure 7: Safety and Security Framework

To what extent do you agree with this measure?

- ☐ The measure should be removed in its entirety
- ☐ The measure should be substantially edited and/or further clarified

X The measure should be lightly edited and/or further clarified

The measure is close to where it needs to be

Please explain your rating to this measure

500 character(s) maximum

It could be clarified whether the SSF is developed for each model or for each model provider.

Please provide suggestions on how to improve this measure

500 character(s) maximum

It could be clarified to whom the SSF is made available.

Measure 12: Mitigations

To what extent do you agree with this measure?

The measure should be removed in its entirety

X The measure should be substantially edited and/or further clarified

The measure should be lightly edited and/or further clarified

The measure is close to where it needs to be

Please provide suggestions on how to improve this measure

500 character(s) maximum

Mitigations should include both proactive and reactive measures. See also our comments at Section Overall Code of Practice. Furthermore, systemic risks should be reduced below an intolerable level, and to *a tolerable level.*

What KPI would you add for this measure?

500 character(s) maximum

See our Consultation free-text input in WG3 for various risk mitigations alongside descriptions, which may be adapted. See also our KPI on whether signatories commit to inform downstream users about autonomous abilities of their models, and the adequate levels of supervision this entails, or "Demonstrating a "margin of safety" for the worst plausible system failures"

Other KPIs:
(1) Diversity of control mechanisms, (2) Resilience to common mode failures, (3) Coverage across attack vectors

Sub-Measure 12.1: Safety mitigations

To what extent do you agree with this sub-measure?

The sub-measure should be removed in its entirety

X The sub-measure should be substantially edited and/or further clarified

The sub-measure should be lightly edited and/or further clarified

The sub-measure is close to where it needs to be

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

We propose that "behavioural modifications" is replaced with "modifications" or a disambiguation, as sometimes modifications can be more structural than behavioral (for example in cases where model internals are understood well enough to modify its structure).

Rephrasing of sentence in Code: mitigations will (replacing "should") be proportional to the risks (Replacing "systemic risk indicators or tiers of severity"), should reduce risks to acceptable levels (added)

What KPI would you add for this sub-measure?

500 character(s) maximum

KPI: Whether and to what extent a model passes various safety benchmarks
KPI: Did the provider red-team the proposed risk mitigations, e.g. testing whether a monitoring technique successfully detects a model crafted by the red team to be faulty or malicious

Sub-Measure 12.2: Security mitigations

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [X] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

We'd like expansion on the possible set of adversaries and threat levels that should be considered by developers, for example whether nation states or other companies are considered possible threats.

What KPI would you add for this sub-measure?

500 character(s) maximum

Did providers of frontier-level GPAI systems establish a red-teaming setup in which well-resourced teams attempt to exfiltrate model-weights?
Did the red-team succeed at their task in the first few rounds of discovering and mitigating issues?

What standards for cybersecurity and information security should be applied to General-Purpose AI Models with Systemic Risk, depending on the systemic risk indicators and tiers of severity? *500 character(s) maximum*

The sub-measure could refer to a comprehensive report by RAND: "Securing AI Model Weights" (https://www.rand.org/content/dam/rand/pubs/research_reports/RRA2800/RRA2849-1/RAND_RRA2849-1.pdf)

In what ways should cybersecurity standards for General-Purpose AI Models with Systemic Risk be different from existing cyber security standards in other domains?

500 character(s) maximum

Stealing frontier GPAI model weights might become the goal of nation state actors, and thus high levels of cybersecurity will be necessary, to a degree unfamiliar with current industry standards.

Sub-Measure 12.3: Limitations

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- x ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

We request clarification on whether the SSF should contain an overview on limitations of mitigations in the industry or in the current processes of the provider, and support the documentation of risks that cannot be mitigated yet.

What KPI would you add for this sub-measure?

500 character(s) maximum

Retrospective validation:
1. Prediction success rate: Number of materialized risks that were documented in the limitations
2. False positive rate: Number of documented limitations without any materialized risks

Sub-Measure 12.4: Process for assessing adequacy of mapping

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ x The sub-measure is close to where it needs to be

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

The measure could state which possible changes in external factors could be considered too implausible to be relevant for the process for mapping continued adequacy.

What KPI would you add for this sub-measure?

500 character(s) maximum

How many contingencies for changed internal or external factors does the process for assessing adequacy consider?

Measure 13: Safety and Security Reports

To what extent do you agree with this measure?

- ☐ The measure should be removed in its entirety
- ☐ The measure should be substantially edited and/or further clarified
- X ☐ The measure should be lightly edited and/or further clarified

☐ The measure is close to where it needs to be

Please provide suggestions on how to improve this measure

500 character(s) maximum

It would be helpful to add a short summarizing sentence on what the contents of the SSR can be soon after it is named, e.g. "This report will contain the results of risk and risk mitigation assessments and shall […]".

Sub-Measure 13.1: Proportionality

To what extent do you agree with this sub-measure?

☐ The sub-measure should be removed in its entirety

☐ The sub-measure should be substantially edited and/or further clarified

X ☐ The sub-measure should be lightly edited and/or further clarified

☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

We ask that terms like "comprehensiveness" and "proportional" be expanded upon.

Sub-Measure 13.2: Results of risk assessment

To what extent do you agree with this sub-measure?

☐ The sub-measure should be removed in its entirety

☐ The sub-measure should be substantially edited and/or further clarified

☐ The sub-measure should be lightly edited and/or further clarified

X ☐ The sub-measure is close to where it needs to be

Sub-Measure 13.3: Results of safety mitigations assessment

To what extent do you agree with this sub-measure?

☐ The sub-measure should be removed in its entirety

☐ The sub-measure should be substantially edited and/or further clarified

☐ The sub-measure should be lightly edited and/or further clarified

X ☐ The sub-measure is close to where it needs to be

Sub-Measure 13.4: Results of security mitigations assessment

To what extent do you agree with this sub-measure?

☐ The sub-measure should be removed in its entirety

☐ The sub-measure should be substantially edited and/or further clarified

☐ The sub-measure should be lightly edited and/or further clarified

X ☐ The sub-measure is close to where it needs to be

Sub-Measure 13.5: Cost-benefit analysis

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- X ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

We consider using cost-benefit analysis as the primary analytical tool to make deployment decisions to be incompatible with both a) best practices in safety engineering and fundamental rights impact analysis and b) the wording of article 55(b) and Article 1 of the AI Act. We refer to point b) in the PREAMBLE.
It could be the case that costs are incurred by the public and benefits are reaped by the providers, which is not sufficient risk mitigation even if benefits outweigh costs.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

We'd appreciate clarifications as to who the relevant parties are that carry the costs and receive the benefits from the deployment.

Sub-Measure 13.6: Sufficient detail on methodology

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- X ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

Generally accepted state-of-the-art risk assessment methodologies incorporate a broader range of reasoning than what is found in 'the scientific method only; therefore saying 'sufficient scientific detail' can be read as giving signatories a free pass not to describe the non-scientific aspects, e.g. related to legal interpretations or fundamental rights impact assessments.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

Remove word 'scientific' in 'Signatories will ensure an SSR has sufficient scientific detail to allow for…' See our attached input 'Feedback on rigour, scientific and other types' for more detail.

The sub-measure could identify which parties would be relevant for independent assessment of methods, e.g. independent auditors, contractors, government agencies.

Sub-Measure 13.7: Review

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- X ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

We'd appreciate explanations of what risks are classed at a high level of severity, and which external parties would perform such a review. It would also be helpful if it was clearer to what degree the internal and external review will be available to government authorities and the public.

Sub-Measure 13.8: Equivalency

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- X ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

The measure is necessary and easy and straightforward to describe.

Measure 14: Development and deployment decisions

To what extent do you agree with this measure?

- ☐ The measure should be removed in its entirety
- ☐ The measure should be substantially edited and/or further clarified
- X ☐ The measure should be lightly edited and/or further clarified
- ☐ The measure is close to where it needs to be

Please explain your rating to this measure

500 character(s) maximum

The general structure of this measure is fine and it captures high level best practices in safety engineering well. Major problem is the reference to 'cost-benefit analysis'.' See our comments under 13.5 and 14.2 for more details.

Please provide suggestions on how to improve this measure

500 character(s) maximum

Measure 14 could be moved ahead of measure 12 – this would make cross-referencing in the draft code more logical and easier to follow for readers. See under 14.2 for more details.

We would also suggest adding deployment practices, such as know-your-customer (KYC) practices and staged release strategies.

What KPI would you add for this measure?

500 character(s) maximum

*KPIs will preferably say how much documentation detail is needed to satisfy '*Signatories will detail in their SSF'

Sub-Measure 14.1: Conditions for not proceeding

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- X ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

The measure should specify more explicitly that GPAI models with systemic risk should not be deployed in situations where systemic risks cannot be reduced below an intolerable level.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

Add sentence: 'One condition for not proceeding with deployment will be that the risk mitigation measures detailed in the SFF are judged by that signatory to be not sufficient to keep post-deployment systemic risks below an intolerable level (see measure 12).'

What KPI would you add for this sub-measure?

500 character(s) maximum

See general answer for 14.

Sub-Measure 14.2: Conditions for proceeding

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- X ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

We consider the use of cost-benefit analysis as the primary analytical tool to make deployment decisions to be

incompatible with both a) best practices in safety engineering and fundamental rights impact analysis and b) the wording of article 55(b) and Article 1 of the AI Act.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

Remove the words 'or through the presentation of a cost-benefit analysis'. More generally, please refer to our comments on 13.5. We prefer that 13.5 is edited and significantly expanded.

What KPI would you add for this sub-measure?

500 character(s) maximum

See general answer for 14.

Sub-Measure 14.3: External input and decision-making

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [x] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

We propose to add that the input and authorisation should be done "to a degree proportional" to the level of risk posed by the GPAI systems.

What KPI would you add for this sub-measure?

500 character(s) maximum

See general answer for 14.

Section IV: [Working Group 4] Governance risk mitigation for providers of General-Purpose AI Models with Systemic Risk

In this section you are asked to provide your overall opinion on the measures and sub-measures included in the fourth section of the Code of Practice related to *governance risk mitigation for providers of General Purpose AI Models with Systemic Risk.*

Measure 15: Systemic risk ownership

To what extent do you agree with this measure?

- [ ] The measure should be removed in its entirety
- [ ] The measure should be substantially edited and/or further clarified

☐ The measure should be lightly edited and/or further clarified

x ☐ The measure is close to where it needs to be

Please provide suggestions on how to improve this measure

500 character(s) maximum

Suggest to also include "Internal independent oversight functions in a transparent governance structure, such as related to risks and ethics" as per initial consultation survey.

Sub-Measure 15.1: Executive level

To what extent do you agree with this sub-measure?

☐ The sub-measure should be removed in its entirety

☐ The sub-measure should be substantially edited and/or further clarified

x ☐ The sub-measure should be lightly edited and/or further clarified

☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

There should be accountable and (optionally) responsible parties assigned for all systemic risks identified.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

Suggest to also include "ensuring that staff are familiar with their duties and the organisation's risk management practices" as per initial consultation survey.

Some examples of what validly counts as being the 'executive level' in SMEs, non-profits, and open source projects should be provided.

What KPI would you add for this sub-measure?

500 character(s) maximum

All systemic risks identified must have at least one person ("accountable party") within the organization who is accountable for.

Sub-Measure 15.2: Board level

To what extent do you agree with this sub-measure?

☐ The sub-measure should be removed in its entirety

x ☐ The sub-measure should be substantially edited and/or further clarified

☐ The sub-measure should be lightly edited and/or further clarified

☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

Some organizations (e.g. SMEs, non-profits) may not have the manpower to maintain a board-level committee separate from an executive-level committee.

Should the above sub-measure be made relative to provider size or other relevant characteristics? If so, how?

500 character(s) maximum

For GPAI model providers who do not have a board, the obligations of the board in this sub-measure should be fulfilled by the executive level, where possible.

Should there be more, or other examples, of what might qualify as adherence to measure 15?

500 character(s) maximum

In the absence of a board-level committee, teams involved in developing models with systemic risk have a clear and direct line of communication with the executive-level. Importantly, Sub-measure 15.1 must be fulfilled.

Measure 16: Adherence and adequacy assessment

To what extent do you agree with this measure?

- ☐ The measure should be removed in its entirety
- X ☐ The measure should be substantially edited and/or further clarified
- ☐ The measure should be lightly edited and/or further clarified
- ☐ The measure is close to where it needs to be

Please explain your rating to this measure

500 character(s) maximum

The SSF as described in Measure 7 does not currently have enough detail to provide feedback on adherence and adequacy. For instance, it is not clear if one SSF is required per model provider or per model.

Please provide suggestions on how to improve this measure

500 character(s) maximum

For ease of compliance, one SSF per model provider may suffice, except for models qualitatively very different in terms of operation, algorithm, or intended use from existing models.

Sub-Measure 16.1: Periodic SSF assessment

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [X] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Are there specific questions such an assessment should answer?

500 character(s) maximum

- Are mitigations scalable to next generation frontier systems, given trends of foreseeable performance improvement?
- Is there sufficient exploration to identify unknown risks?

How should adequacy be defined in this context?

500 character(s) maximum

Adequacy can mean the following: (1) adherence to the SSF is clear and understandable to all stakeholders inside the organization (rather than being fuzzy or vague) (2) failures in adherence are easy to detect and correct (either internally or by AI Office), (3) that adherence to the SSF sufficiently mitigates the current and foreseeable risks, (4) the SSF is somewhat consistent with other similar GPAI providers in its level of rigor, and (5) the SSF fully meets the requirements laid out in the Code.

Measure 17: Independent expert systemic risk and mitigation assessments

To what extent do you agree with this measure?

- [ ] The measure should be removed in its entirety
- [X] The measure should be substantially edited and/or further clarified
- [ ] The measure should be lightly edited and/or further clarified
- [ ] The measure is close to where it needs to be

Please explain your rating to this measure

500 character(s) maximum

It appears that independent expert systemic risk and mitigation assessments are optional ("as appropriate"). However, the question of when to involve independent experts in the model lifecycle is not specified, which risks this sub-measure becoming moot.

Please provide suggestions on how to improve this measure

500 character(s) maximum

Clear criteria for requiring the involvement of independent experts to ensure that they are conducted as appropriate.

Under what circumstances is independent expert systemic risk assessment of a General-Purpose AI Model with Systemic Risk appropriate before deployment? What about assessment of mitigations? Under what conditions does it seem counterproductive or unnecessary?

500 character(s) maximum

Independent expert risk assessment is crucial when in-house assessment may not be sufficiently representative, due to: 1) The provider lacking in-house expertise (e.g., CBRN testing), 2) significant advance in capabilities / novel training approach, carrying risks the provider may not be equipped to assess 3) Other

factors indicating in-house assessments may not fully capture model risks. We recommend further plenary discussions to clarify conditions where it could be made mandatory.

Are there circumstances under which it is appropriate or advisable to involve independent experts in risk assessments iteratively, throughout the lifecycle, starting before or during training? *500 character(s) maximum*

Stronger evaluations and assessment are advisable when the risk profile of a model is expected to be higher. This can happen when there are indications that a model may have novel risks, e.g., when it is trained on more data or compute than previous state-of-the-art models, or other models showing frontier capabilities despite not necessarily being trained on more data or compute, but by other means e.g., by using a different training setup (e.g. o1) or scaffold (e.g. a new agent scaffold).

How can independent systemic risk assessments be adapted to the magnitude and nature of the relevant systemic risk, e.g. with regards to inform security, depth of access to General-Purpose AI Models with Systemic Risk components and documentation, scope of testing, time to test, expertise and transparency? *500 character(s) maximum*

The SSF should detail the risk identification, assessment, and mitigation required for models with different levels of systemic risks. If a model has a sufficiently high level of systemic risk, it may require more stringent evaluation and assessment, which may require the involvement of independent experts, among other implications.

Sub-Measure 17.1: Before deployment

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- ☐ The sub-measure should be lightly edited and/or further clarified
- X ☐ The sub-measure is close to where it needs to be

What constitutes an appropriate third-party evaluator? How can the Code of Practice be drafted so as to take into account the current immaturity of the industry? Is there some way providers, especially Small Medium Enterprises (SMEs), can be supported by the AI Office in ensuring independent expert assessment of risks and mitigations?

500 character(s) maximum

Independent experts should fulfill the following criteria: 1) They do not have financial or personal ties to the model provider being evaluated or assessed, 2) They have significant expertise in the areas in which the evaluation or assessments are being performed, 3) Their methodology of assessment is transparent and follows best practices, and 4) They provide clear reporting of findings and make actionable recommendations, including recommendations on model deployment decisions.

Sub-Measure 17.2: After deployment

To what extent do you agree with this sub-measure?

- ☐ The sub-measure should be removed in its entirety
- ☐ The sub-measure should be substantially edited and/or further clarified
- X ☐ The sub-measure should be lightly edited and/or further clarified
- ☐ The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

We appreciate the inclusion of this Sub-Measure. It may still benefit from further detail or sub measures to reduce ambiguity for providers, especially in terms of what it means to "enable meaningful independent testing", specifically on what constitutes "sufficient access" to "independent researchers" and "other relevant parties", especially given potential concerns on trade secrets.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

Add "Responsible Disclosure" - similar to cyber, companies should communicate norms on vulnerability disclosure (e.g. a novel jailbreak) or set up channels for private reporting in an organized way, with sufficient time to patch the vulnerability before public disclosure. It should also inform other labs (or AI Office), in the case where a vulnerability is suspected to be transferable across models.

Providers should take care to avoid abuse of researcher access, which can be a security risk.

When are different means of facilitating independent testing – such as research safe harbors and vulnerability reporting – appropriate?

500 character(s) maximum

Good faith security research / research safe harbor should be encouraged, as long as it can be distinguished from bad actors making use of the models. For example:
-Academic research
-Long-term impact studies
-Emerging risk evaluation

Vulnerability reporting should be used for model security and robustness flaws.

Different access levels should be used, depending on research. Some research will require internal model access (e.g. linear probes), while others can be API access.

Measure 18: Serious incident reporting

To what extent do you agree with this measure?

- [ ] The measure should be removed in its entirety
- [ ] The measure should be substantially edited and/or further clarified
- [x] The measure should be lightly edited and/or further clarified
- [ ] The measure is close to where it needs to be

Please explain your rating to this measure

500 character(s) maximum

AI Act recital 49:
'serious incident' means an incident or malfunctioning of an AI system that...leads to any of the following:
- the death of a person, or serious harm
- a serious…disruption of…critical infrastructure;
- the infringement of…fundamental rights;
- serious harm to property or the environment;
There should be further clarity on what constitutes a serious incident, e.g. what does "serious harm to a

person's health" or "serious harm to property or the environment" entail

What KPI would you add for this measure?

500 character(s) maximum

Possibly relevant, from Aviation/Nuclear:
Incident Management Effectiveness Score
-Detection speed
-Classification accuracy
-Response time
-Investigation quality
-Corrective action effectiveness

What does a serious incident entail? Should the Code of Practice use the definition the AI Act uses for AI systemic in Article 3(49) or is another definition more appropriate for General-Purpose AI Models with Systemic Risk?

500 character(s) maximum

The Code should stay within the AI Act definition for AI Systems and merely clarify it where it is left open ended, with the help of examples that better represent GPAI System capabilities and risks.

Under what conditions should a General-Purpose AI Model with Systemic Risk be judged to have indirectly led to a serious incident occurring?

500 character(s) maximum

Some factors:
-The model's capabilities or known vulnerabilities were an enabling factor in the incident, even if not the sole cause
-The incident occurred through downstream AI systems incorporating the model
-The risk was inadequately mitigated despite being identified in the SSF

Are there suitable technical standards or best practices that can be enable automated or streamlined reporting of serious incidents to the AI Office?

500 character(s) maximum

(1) Possibly AIS (Automated Indicator Sharing) as per CISA, e.g. using STIX/TAXII.
(2) APIs for automated submission to AI office
(3) Templates for incident metadata and classification taxonomy

Sub-Measure 18.1: Serious incident reporting processes

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [X] The sub-measure is close to where it needs to be

Sub-Measure 18.2: Response readiness

To what extent do you agree with this sub-measure?

X The sub-measure should be removed in its entirety

The sub-measure should be substantially edited and/or further clarified

The sub-measure should be lightly edited and/or further clarified

The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

We agree that it is crucial for response readiness to be included in the Code, but we believe it should not fall under "serious incident reporting" as it is a matter of (reactive) risk management unrelated to incident reporting.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

Add sub-measure 18.2 as part of the specification of the SSF. Also, sub-measure 18.1 should refer to the response readiness outlined in the SSF, in particular what measures were taken when the serious incident occurred.

What possible corrective measures could be taken in response to serious incidents? Should the Code of Practice specify when they may be appropriate?

A few examples: Model recall or shutdown, user access restriction, model sandboxing, information sharing if an incident may be due to an adversarial attack or vulnerability which is transferable to other models (e.g, with similar architecture or training data).

What serious incident response processes are appropriate for open weight or open-source providers?

Norms in cybersecurity can be followed, with community vulnerability reporting, alerts, documentation updates etc.

Measure 19: Whistleblowing protections

To what extent do you agree with this measure?

The measure should be removed in its entirety

The measure should be substantially edited and/or further clarified

X The measure should be lightly edited and/or further clarified

The measure is close to where it needs to be

Please explain your rating to this measure

500 character(s) maximum

Since the directive is not directly enforceable for most AI labs, it should be reiterated or explicitly detailed here. For example, legal protections for employees disclosing sensitive information, intellectual property, or trade secrets, provided they exercise reasonable care to disclose only details relevant to the violation.

Please provide suggestions on how to improve this measure

500 character(s) maximum

Further assurance to whistleblowers, e.g., 1) The AI Office will not notify the employer when this happens, and protect employee identity, 2) The complaints depository will be reasonably secure,

Concrete guidance for the whistleblower, e.g. the nature of the complaint should be either 1) Company is violating the law 2) Company's actions contradict their SSF, etc. 3) Company is making unsafe decisions, due to inadequate SSF or other reasons (e.g. internally deploying a powerful model)

What KPI would you add for this measure?

500 character(s) maximum

Are independent retaliation monitoring or anti-retaliation systems in place (to resolve allegations of retaliation)?
Does the organization culture incentivise speaking up, and encourages fair resolution of issues?
Are employees and managers properly informed of their rights and obligations?
Independent audit to determine if the program is working.

Sub-Measure 19.1: Inform

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [x] The sub-measure should be lightly edited and/or further clarified
- [ ] The sub-measure is close to where it needs to be

Please explain your rating to this sub-measure

500 character(s) maximum

The AI Office mailbox is not adequately described.

Please provide suggestions on how to improve this sub-measure

500 character(s) maximum

The AI Office mailbox should be reasonably secured and afforded dedicated personnel to review and respond to any incoming whistleblowing and notifications within reasonable time.

Measure 20: Notifications

To what extent do you agree with this measure?

- [ ] The measure should be removed in its entirety
- [ ] The measure should be substantially edited and/or further clarified
- [ ] The measure should be lightly edited and/or further clarified
- [x] The measure is close to where it needs to be

Sub-Measure 20.1: General-purpose AI model with systemic risk notification

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [x] The sub-measure is close to where it needs to be

The AI Office has the authority to update the classification criteria for determining whether a General Purpose AI Model is presumed to have high-impact capabilities (and are therefore whether it is classified as a General-Purpose AI Model with Systemic Risk). How could it be written such that it is clear when providers should notify the AI Office of a model meeting new classification criteria? *500 character(s) maximum*

In the case of systemic risks of GPAI systems, a clear criteria may be challenging. It may be helpful to encourage the provider to consult with the AI office, instead of trying to create a very unambiguous criteria that may become an imperfect proxy over time.

Sub-Measure 20.2: SSF notification

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [x] The sub-measure is close to where it needs to be

Sub-Measure 20.3: SSR notification

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [x] The sub-measure is close to where it needs to be

Sub-Measure 20.4: Substantial systemic risk notification

To what extent do you agree with this sub-measure?

- [ ] The sub-measure should be removed in its entirety
- [ ] The sub-measure should be substantially edited and/or further clarified
- [ ] The sub-measure should be lightly edited and/or further clarified
- [x] The sub-measure is close to where it needs to be

What constitutes a strong reason to believe systemic risk might materialise?

500 character(s) maximum

Some examples may include 1) Model evaluation results showing capabilities that exceed SSF defined thresholds, 2) Unexpected capabilities during internal testing, 3) Identification of vulnerabilities 4) Near misses, incidents, or trends suggesting potential future escalation, 5) Other considerations, e.g. immediacy, rapid progress on risk indicators, narrow window for intervention, risks that may arise quickly or with less warning or high severity may warrant a lower bar for "strong reason".

Measure 21: Documentation

To what extent do you agree with this measure?

- [ ] The measure should be removed in its entirety
- [ ] The measure should be substantially edited and/or further clarified
- [X] The measure should be lightly edited and/or further clarified
- [ ] The measure is close to where it needs to be

Please explain your rating to this measure

500 character(s) maximum

It is not clear if documentation, when submitted,is required at a regular interval to ensure that updated documentation is provided to the AI Office. This is because the Measure mentions that sharing of information to the AI Office is to be done only "upon request".

What could a standardised template for such documentation look like, to reduce compliance costs, especially for smaller providers? Note in future drafts, we intend to ensure the documentation under this measure is streamlined and combined other documentation requirements such as those detailed in Annex XI, Section 1, and Annex XII.

500 character(s) maximum

1.Basic Information:
- Provider details
- Model identification
- Version information
2.Risk Assessment:
- Risk identification
- Mitigation measures
- Monitoring plans
- Response procedures
3.Technical Details:
- Training data summary (from Measure 1)
- Model architecture
- Known/ projected Capabilities
- Limitations
- Safety features
- Security measures
4.Compliance Evidence:
- Control implementation
- Testing results
- Audit findings
- Incident reports
- Update history

Measure 22: Public transparency

To what extent do you agree with this measure?

- [ ] The measure should be removed in its entirety
- [ ] The measure should be substantially edited and/or further clarified

X ☐ The measure should be lightly edited and/or further clarified

☐ The measure is close to where it needs to be

Please explain your rating to this measure

500 character(s) maximum

All information submitted to the AI Office should be made available to the public, with the exception of sensitive commercial information, in which case the AI Office should be informed accordingly (i.e. some information is not disclosed to the public).

For what types and levels of public transparency do systemic risks increase, instead of decreasing by empowering the broader ecosystem to assess and mitigate them?

500 character(s) maximum

1) disclosing technical details, especially without context (e.g. releasing only open-weights), can enable malicious actors to exploit model vulnerabilities or capabilities

2) Providers, especially outside the EU market, may use the transparency requirement to accelerate development of capabilities while not being compelled to disclose. Some disclosure may be best saved for once there is wider international agreement on safety practices, rather than just companies entering the EU market.

How burdensome is this kind of public transparency, given the common practice of publishing model and system cards? Can the measure be designed to reduce such burdens?

500 character(s) maximum

A significant burden can be deciding which data to include in reporting, aside from the data from model and system cards that should be included in the SSF and SSRs by default. To reduce the burdens on the provider, the Office should standardize the SSF and SSR with a template that can handle providers of varying sizes. The Office should also specify where to provide such reports (e.g. a dedicated page on the provider's website).

Supporting Documents

See supporting documents (which were attached to this survey submission) on our website.

For any questions or comments, please email inquiries@aistandardslab.org