HUGGING FACE

Third Draft of the General-Purpose AI Code of Practice Feedback Surveys

Working Group 1 & 2

Context:
https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts

WG 1 Transparency and copyright-related rules

Transparency Section Preamble

Please provide below your thoughts on the preamble of the Transparency Section of the Code of Practice.

We welcome thorough transparency requirements, as they ensure safety and security for both downstream providers as well as the users of deployed AI systems. Providing a transparency template constitutes a welcome contribution toward making those requirements more efficient in a way that works for small and large developers alike.

We also appreciate the inclusion of language in the Code of Practice to reflect Article 53(2)'s exemption for models under FOSS licenses.

We are however somewhat concerned to see that several categories of information on non-FOSS GPAI that are necessary for downstream providers now have their disclosure limited to the National Competent Authorities within the Model Documentation Form, which in our view leaves some significant gaps for meeting the requirements set out in Article 53(1.b.i) that the documentation provided "enable providers of AI systems to have a good understanding of the capabilities and limitations" of GPAI and have sufficient information to comply with their own obligations - including notably Articles (9), (10), and (15) where the GPAI is a component of a high-risk system.

Of particular concern with regards to those requirements are the lack of sufficient information about the composition of the GPAI training data to meaningfully assess their suitability for specific (potentially high-risk) applications or to enable sufficient in-context evaluation without undue risk of benchmark contamination, and the lack of sufficient

information about model evaluations to assess their relevance to a particular deployment setting. Additionally, specific information about measures taken to address prohibited or personal data during training are necessary for building a robust risk management system for a high-risk application leveraging the GPAI (Article (9.2)).

While we understand that this information is meant to be covered in "Signatories commit to providing additional information necessary to enable downstream providers" – with the Model Documentation Form more focused on Article 53(1.b.ii) – we do wish to express significant concerns that, without further standardization or accountability mechanisms, having to seek that information on a case-by-case basis puts downstream providers at a significant disadvantage in situations where they may be directly competing with the providers of GPAI in developing commercial AI products for high-risk settings. In particular, any information provided by Downstream Providers to justify the need for specific information will likely include information about market fit studies or product design that can easily be abused.

In order to address these concerns, and given the similarities between the information needed for NCAs and the AI Office and the information needed to fulfill regulatory requirements for providers of high-risk AI systems leveraging GPAIs (see our further response below), we propose leveraging the information contained in the Model Documentation Form directly to additionally meet the requirements of Article 53(1.b.i) for Downstream Providers with a legitimate project of developing an AI system in a high-risk setting. This could take the form, for example, of having the AI Office or NCAs examine requests from Downstream Providers applying for access to the information they need for a given use case, and in case of legitimate requests sharing the relevant subsets of the Model Documentation Form on training and evaluation subject to the Confidentiality Requirements set out in Article 78 - additionally protecting information submitted by Downstream Providers.

Overall, a stronger commitment to public disclosure, even within the already narrow set of transparency categories, would do far more to support public safety and governance, innovation, and fair competition.

Model Documentation Template: Transparency

Transparency, Item 6: Information on the data used for training, testing, and validation

To what extent do you agree with this item?

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

HUGGING FACE

Please explain your rating and suggest improvements, clearly distinguishing between concerns regarding scope/level of the item and drafting improvements

Currently, the information provided to Downstream Providers about training, testing and validation is limited. This generally stands in the way of their ability to build robust and reliable technology in general, and precludes Downstream Providers working on AI systems considered high-risk from meeting their obligations under the AI Act.

Downstream Providers, especially ones working on high-risk AI systems, should have access to the following three additional sources of information to understand whether the model they are deploying could potentially perpetuate harm:

Measures to detect unsuitability of data sources (harmful data)
Measures to detect unsuitability of data sources (personal data)
Measures to detect identifiable biases

We argue specifically that all of this information is necessary for downstream providers to meet their commercial and regulatory requirements following Article 53(1.b). Evaluation of a model in context is a core requirement of robust technology, but is currently threatened by endemic issues of data contamination between training data and in-context evaluation benchmarks.[1] Published research has also consistently shown that pre-training interventions alone on privacy and harmful content is insufficient, and that downstream developers need to be sufficiently informed to develop complementary mitigations.[2][3] For developers of high-risk AI systems, information about the training data overall composition and some level of detail on training data processing and harm mitigation is therefore necessary to meeting the requirements of Articles 9(2.a), 9(2.b), and 9(2.c) on in-context risk identification and management, and Article 10(2.e) and 10(2.c) and assessing the suitability for their use cases and use-case relevant biases.

Additionally, we note that Annex XII (2.c) requires information about “data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies”. In the current template, only training data is documented in any level of detail, but we would strongly encourage the chairs to read the second part of this statement as inclusive of the testing and validation data, as we have seen that minor differences in the processing of benchmark data can lead to drastically different results.[4][5] For developers of high-risk AI systems, this information will also be critical for meeting the requirement of Article 15(1).

[1] NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark - https://aclanthology.org/2023.findings-emnlp.722/
[2] From Representational Harms to Quality-of-Service Harms[...] - https://aclanthology.org/2024.findings-acl.927/
[3] Upstream Mitigation Is Not All You Need[...] -

https://aclanthology.org/2022.acl-long.247/
[4] Fixing Open LLM Leaderboard with Math-Verify -
https://hf.co/blog/math_verify_leaderboard
[5] Let's talk about LLM evaluation -
https://huggingface.co/blog/clefourrier/llm-evaluation

Transparency, Item 8: Energy consumption

To what extent do you agree with this item?

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

Please explain your rating and suggest improvements, clearly distinguishing between concerns regarding scope/level of the item and drafting improvements

Information about the energy consumption of GPAI is valuable to downstream providers, including for providers who want to manage the energy footprint of their entire stack and to reasonably forecast financial cost evolution for models their commercial offerings rely on.

Transparency, Item 9: Additional information to be provided by providers of general-purpose AI models with systemic risk

To what extent do you agree with this item?

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

Please explain your rating and suggest improvements, clearly distinguishing between concerns regarding scope/level of the item and drafting improvements

Similar to our points above, the information about the results of evaluation, adversarial testing, and model adaptations need to be available to downstream providers to ensure that they can safely deploy the AI models in a real-world context and avoid known risks of the models.

Commitments by providers of general-purpose AI models: Copyright

Copyright, Commitment I.2: Copyright policy

While we appreciate the effort to provide a phrasing that further specifies the scope of the Copyright policy to "address all commitments pursuant to this Section", we are concerned that this second sentence can still be read as constituting a subset rather than the entirety of the policy. In our view, providing a template for the policy would both clarify this scope and make it much easier for external stakeholders to leverage by moving towards greater standardization as to how the information is presented. We stress also that such a template would be particularly valuable to SMEs and smaller organizations who do not have the extensive legal resources necessary to sufficiently analyze the copyright implications of a fast moving technology.

Copyright, Measure I.2.1: Draw up, keep up-to-date and implement a copyright policy

As stated above, especially smaller developers without a dedicated legal team or even lacking any legal support, would need more guidance on how to draw up such a copyright. More guidance, in the form of an example policy or more direct examples of the minimum required content covered, would support developers with compliance.

Copyright, Measure I.2.2: Reproduce and extract only lawfully accessible copyright-protected content when crawling the World Wide Web

We appreciate the measure addressing concerns of rights holders with regards to the origin of training data w.r.t. paywalls. Particularly the list of web domains that will be provided by the EU will be very impactful for smaller providers that need guidance on which websites they can and cannot scrape. A machine-readable format for this list would be highly appreciated.

Copyright, Measure I.2.3: Identify and comply with rights reservations when crawling the World Wide Web

We appreciate the approach (1b) of best efforts of extending machine-readable protocols further than just robots.txt focuses on “widely adopted” protocols.

We would appreciate clarity on this measure about which information should be reported as part of which process. For example, the information about web crawlers (4) should be accessible to rights holders, but it is not clear if it is the best space here or if it should be part of the AI Office’s training data template and/or part of other transparency

requirements.

Copyright, Measure I.2.5: Mitigate the risk of production of copyright-infringing output

We acknowledge and appreciate the necessary open source exemption for submeasure (1)(b) in submeasure (2).

Copyright, Measure I.2.6: Designate a point of contact and enable the lodging of complaints

Open source projects and projects developed outside of large commercial entities often remain available and valuable for periods that exceed the administrative and legal capacity of their initial developers. For open-source projects which may no longer be actively maintained but remain valuable to the AI developer community, the code should consider realistic and practical solutions for sustaining compliance without imposing excessive burdens. For example, for projects or models that are no longer actively maintained, introduce a sunset clause to limit the obligations of contributors, ensuring that the requirements do not outlast the involvement and availability of the developers.

We appreciate that the current phrasing points out that rightsholders requests should be appropriately substantiated and limited to questions of compliance with the policy outlined in this question. We are however still concerned that the current approach still risks requiring SMEs to have to sort through large amounts of requests and make potentially difficult judgments about whether a third party would find the request sufficiently substantiated.

To address this limitation, we reiterate our recommendation to provide a modular template for the Copyright policy proposed in this Section. We also recommend that sufficiently detailed public documentation of the steps taken to fulfill commitments in this Section should exempt SMEs and smaller developers from answering all queries. This would both alleviate the burden on the less resourced developers and provide valuable transparency to external stakeholders.

WG 2 Risk identification and assessment for systemic risk

Safety and Security Section Preamble

Please provide below your thoughts on the preamble of the Safety and Security Section of the Code of Practice.

While the overall text of this section shows significant progress compared to previous drafts, there is still significant room for growth to ensure that it follows previous commitments co-signed by the AI Office to ensure that risk assessment is guided by scientific principles – including by making it sufficiently transparent, multistakeholder, and reproducible.[1]

One major point of tension seems to exist between having strong evidence and scientific consensus standards for assessing novel risks on the one hand and developers' common arguments that the technology moves "too fast" to be subjected to robust regulatory science with sufficient access to independent evaluators on the other. This is underlined for example by the inclusion of some speculative risk categories in Annex 1.1 ("Loss of Control"), several references to "forecasting" as a method throughout, and text that currently discourages developers from sharing evaluation data that is critical to ensuring robust independent review of their framings, findings, and risk prioritization (II.4.10).

While the cadence of product releases from developers of the systems that are expected to be classified as GPAISRs is certainly impressive, it is important to keep in mind that the risk profiles described by the developers' own system cards have not changed drastically in the year or more since the frameworks were introduced.[2][3] This strongly suggests that there is indeed time to subject risk evaluation methodologies to peer review from independent researchers to ensure that risk management approaches do address the most pressing needs of external stakeholders, enabling sufficiently informed external oversight to play its critical role in supporting robust governance and risk management for technical systems.[4]

In order to balance developers' need for privacy and intellectual property of their technical contributions for GPAI development, we suggest relying on the AI Office's scientific panel of independent experts [5] as a primary resource in identifying systemic risk categories, measures, and mitigations. By limiting the focus of the current code on speculative risk categories and forecasting methods with limited grounding in realized harms and instead prioritizing disclosure of current risk assessment methodologies and relevant information about the systems tested to the scientific panel – or additional experts mandated by the AI Office in collaboration with the scientific panel – the code could be made significantly more future-proof and efficient. Organizations like the EMA and its handling of Company Confidential Information can provide some guidance as to how this approach can meet the confidentiality requirements of Article 78.[6]

[1] https://digital-strategy.ec.europa.eu/en/news/fostering-global-ai-safety-eu-ai-office-participates-inaugural-international-network-ai-safety
[2] https://www.anthropic.com/news/anthropics-responsible-scaling-policy (September

2023)
[3] https://openai.com/safety/I (Preparedness framework released in beta in December 2023)
[4] https://dl.acm.org/doi/10.1145/3514094.3534181
[5] https://digital-strategy.ec.europa.eu/en/policies/ai-office
[6] https://www.ema.europa.eu/en/about-us/how-we-work/transparency

Risk assessment for providers of GPAISRs

Risk assessment, Commitment II.2: Systemic risk assessment and mitigation along the entire model lifecycle, including during model development

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

The clarity and technical feasibility of this measure highly depends on the definition of the systemic risks. As of now, the systemic risks are not defined in a way that would make clear how and which evaluation metrics could be applied. The risks in Appendix 1.1 are not sufficiently detailed or based in science to be practically applicable (CBRN https://arxiv.org/abs/2412.01946; “harmful manipulation” is underdefined; “loss of control” spans both immediate concerns that are already covered by cybersecurity and speculative risks grounded in excessive anthropomorphization of the systems).

Further, while there are extensive details as to what should be recorded in submeasure (4), it is not clear where or to whom this information should be reported. However, the audience of the report would be crucial to understand in which format the required information should be documented.

Risk assessment, Commitment II.3: Systemic risk identification

To what extent do you agree with this commitment?

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

- ☐ The commitment should be lightly edited and/or further clarified
- ☐ The commitment is close to where it needs to be

Please explain your rating and suggest improvements

While the changes in this measure appear to provide a more flexible approach to the systemic risks framework, this measure is still anchored in the list of systemic risks in Appendix 1. We have previously detailed while the risks listed here cannot build a foundation for the systemic risk framework.[1][2] Systemic risks in the AI Act are broadly defined as risks to public health, safety, society as a whole, or to entire domains of activity or communities arising from “high-impact capabilities”. Recital 110 specifically mentions risks tied to “major accidents” or “disruptions of critical sectors” as well as effects on democratic processes and information ecosystems. This brings to mind risks like the recent global outage due to the CrowdStrike failure, or major threats to e.g. access of EU citizens to essential services. These categories of risk arise from “high-impact capabilities”; i.e. systems that are marketed as capable of producing secure software code for critical systems, or as capable of summarizing an individual’s entire history and legal context well enough to assign unemployment or other benefits. They are supported by enough substantive evidence across fields of technology and social sciences to make them a significant and immediate concern. They can also be mitigated individually and collaboratively by signatories of the Code of Practice through design choices, robust and transparent documentation of the model’s performance and limitations for different categories of capabilities, and sufficient access to downstream developers to support in-context evaluation and mitigation approaches.

Some of these risks are represented in the risks listed in Appendix 1.2. However, in this measure these concrete systemic risks are presented as optional in (1) “where it can be reasonably foreseen" linking to Appendix 1.2, whereas the risks in Appendix 1.1 are non-negotiable. This undermines the validity of the risk assessment. Risks in Appendix 1.2 should be part of the core definition of systemic risks, better aligning with the intention of Recital 110.

By contrast, risks that are significantly less likely to be realized in the near future if at all, such as loss of human control from increased capabilities, can be incorporated at a later stage, once they become relevant and appropriate evaluation methodologies have been developed, as enabled by Annex 2.

[1] https://huggingface.co/blog/yjernite/eu-draft-cop-risks
[2] https://huggingface.co/blog/frimelle/eu-third-cop-draft

Risk assessment, Measure II.3.1: Systemic risk selection

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

This measure should be edited from two points:
1. As detailed above, the measures listed in Appendix 1.1 and 1.2 should be reorganised to better reflect the intentions of Recital 110. Accordingly, reporting on the risks detailed on Appendix 1.2 should not be presented as optional. While we have reservations with some of the risks listed in Appendix 1.2 (see our comments on the Appendix 1.2 below), they align better with the current risks observed regarding AI models.
2. This measure's focus shift towards centering submeasure (2). Submeasure allows for more flexibility in the risk identification, therefore future-proofing the code of practice and the covered risks.

Risk assessment, Measure II.3.2: Determining systemic risk scenarios

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

Previous work has shown that predictions about future developments of AI are erroneous.[1] Rather than attempting to anticipate all possible risk scenarios in advance, it would be more effective to prioritize comprehensive documentation and empirical evaluation of known risks. Such an approach would enable a broader group of stakeholders to contribute to identifying and assessing risks, rather than relying solely on developers' internal assumptions.

At this stage, it remains unclear how this measure aligns with the requirements of the EU AI Act, or what would constitute sufficient compliance, e.g., how many risk scenarios are expected, or how detailed they should be. Without a scientific standard for developing

such scenarios, there is a risk they may not reflect realistic or evidence-based concerns.

[1] https://www.fhi.ox.ac.uk/wp-content/uploads/FAIC.pdf

Risk assessment, Commitment II.4: Systemic risk analysis

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

Overall, the measure presents reasonable risk analysis techniques, which are based in scientific and state of the art approaches to measuring risk. However, to the link to list of systemic risks in Appendix 1.1, these measures cannot be taken with reasonable effort, as these risks are open for interpretation.

Risk assessment, Measure II.4.1: Systemic risk estimation

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

The submeasure adequately describes a methodology to evaluate and record risks for an AI model. If the definition of systemic risk focuses on measurable, real risks of AI models, this measure summarises (in great detail) a good methodology to record and report the evaluation results with a focus on risk. However, estimations without a robust methodology such as “(3) a quantitative estimation of the likelihood and expected harm (e.g. “our model has an X% likelihood of causing €Y in losses due to systemic risk Z”).” should not be an option besides qualitative, measurable, and provable risk factors.

Risk assessment, Measure II.4.4: State-of-the-art model evaluations

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

We would appreciate the addition of “non-SME Signatories commit to actively striving to also improve on publicly available methods.” as in the second draft Measure 10.2. State-of-the-art model evaluations - this allows for wider sharing, the improvement of evaluation and encourages updating of the state of the art.

Risk assessment, Measure II.4.5: Rigorous model evaluations

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

We appreciate the acknowledgement that SMEs might not have the resources in-house and are able to find a suitable format for reporting in agreement with the AI Office.

Risk assessment, Measure II.4.6: Model elicitation

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

The current draft is here very prescriptive and could benefit from moving additional details into guidelines rather than core parts of the measure, as is here the case for the entire measure. It could be one possible guideline for the model evaluation measure.

Risk assessment, Measure II.4.7: Models as part of systems

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

In the case of open source AI models, model providers cannot monitor the use of their model, nor can their distribution be limited by collecting information on downstream usage, as it would limit the possibility of distributing a model under a FOSS license. Here, open source models need to be exempt from collecting of usage data.

Risk assessment, Measure II.4.9: Exploratory research and open-ended red-teaming

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

In submeasure (1), listing “research in support of forecasting” among scientific evaluation methodologies dilutes the purpose of the evaluation methodologies. Forecasting is not a valid scientific evaluation methodology, see our answer to Risk assessment, Measure II.4.13: Evaluation practices for forecasting.

Risk assessment, Measure II.4.10: Sharing tools & best practices

To what extent do you agree with this commitment?

- ☐ The commitment should be removed in its entirety

- ☑ The commitment should be substantially edited and/or further clarified
- ☐ The commitment should be lightly edited and/or further clarified
- ☐ The commitment is close to where it needs to be

Please explain your rating and suggest improvements

Whereas the previous draft (2nd draft, Measure 10) explicitly encouraged sharing of evaluations, specifically by non-SMEs, this measure limits Signatories in sharing data, which should be in their decision rather than codified in the CoP; especially the requirement for dedicated staff for sharing research data disincentives a best practice already in use and beneficial to improve the evaluation ecosystem. We suggest rephrasing the recommendation to encourage the sharing of evaluation methods and data "unless there are strong concerns" that the risks outlined in this measure may materialise. First and second party auditing has been shown to be flawed, emphasising the importance of a broader evaluation ecosystem.[1][2] Where public sharing of this information should not be possible for any reason, there are alternative ways of enabling access of third party evaluators to, e.g., training data without having to share the entire dataset.[3]
It is in the interest of the code of practice and the AI Office to incentivise the sharing of evaluation methodology, benchmarks, and all other artefacts, to ensure access to best practices and an overall improved evaluation ecosystem.

[1] https://arxiv.org/abs/2401.14462
[2] https://arxiv.org/abs/2310.02521
[3] https://openmined.org/blog/third-party-evaluation-to-identify-risks-in-llms-training-data/

Risk assessment, Measure II.4.11: Qualified model evaluation teams and adequate evaluation access and resources

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

The description of the necessary evaluation teams is unlikely to be achievable internally by SMEs, both in terms of the required team size and the necessary in-house expertise.

More proportionate expectations should depend on a company's or open-source project's size and available resources. Furthermore, a separate evaluation team may not be realistic or even the best choice in all contexts. An evaluation team embedded within the research team, or having the same person develop and evaluate the model, is a more realistic setup in most smaller companies and teams. This arrangement also allows for seamless communication about model capabilities and potential deployments.

Risk assessment, Measure II.4.12: Safety margins

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

This measure contains more details on how to conduct the evaluations. For readability and ease of application, all measures defining how to run evaluations should be under one measure, giving an easier overview of what requirements should be fulfilled to run adequate evaluations that satisfy compliance.

Risk assessment, Measure II.4.13: Evaluation practices for forecasting

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

Forecasting future AI model capabilities and risks is inherently speculative and should not be treated as a scientific evaluation method. Scientific model evaluation relies on empirical, testable evidence, whereas capability forecasting leaps beyond available data into conjecture. Decades of AI research are replete with confident forecasts that proved wrong. Analyses of past predictions find they were "all over the map, with no pattern of convergence, and no visible difference between expert and non-expert". [1] Expert opinion has often fared no better than chance in predicting AI breakthroughs. This track record

shows that speculative foresight cannot be treated as a scientific assessment method.

Predictive claims about future AI capabilities cannot be verified or falsified at the time they are made, violating basic scientific criteria and leaving such forecasts outside rigorous evaluation. For example, the notion that ever-larger models will develop qualitatively new abilities is “naturally un-falsifiable”. [2]

Given improved model capabilities are achieved with new data scales and types and through design choices,[3][4] it is more practical to focus on model-agnostic evaluation methods that assess performance across various architectures and training regimes, rather than attempting to predict the evolution of specific model types or the emergence of new capabilities.

[1] https://www.fhi.ox.ac.uk/wp-content/uploads/FAIC.pdf
[2] https://philsci-archive.pitt.edu/23622/1/psa_scaling_hypothesis_manuscript.pdf
[3] https://arxiv.org/abs/2001.08361
[4] https://openaccess.thecvf.com/content/CVPR2022/html/Zhai_Scaling_Vision_Transformers_CVPR_2022_paper

Risk assessment, Measure II.4.14: Post-market monitoring

To what extent do you agree with this commitment?

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

Please explain your rating and suggest improvements

In the case of open source models, post-market monitoring cannot be prescribed if the model provider is not also the model deployer. In itself, the FOSS licenses do not allow for such clauses to allow model providers to monitor downstream usage of their models. Open source models should be exempt, as, e.g., community-driven evaluation happens organically and not influenced by the model provider.

HUGGING FACE

Appendix 1. Systemic Risk Taxonomy

Taxonomy of Systemic Risks, Appendix 1.1: Selected types of systemic risk

To what extent do you agree with this appendix?

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

Please explain your rating and suggest improvements

The systemic risk taxonomy in Appendix 1.1 should either be removed from the Code or substantially amended. Several of the selected categories are speculative, unevenly defined, and not grounded in current scientific consensus. This risks directing disproportionate compliance efforts toward remote hazards while diverting attention from more immediate, evidence-based risks such as AI system failures in critical infrastructure or unintended harms arising from scaled deployment.

In particular, we strongly recommend removing “Loss of Control” as a category. We acknowledge that some of the examples currently in the category correspond to possible negative outcomes of inadequate development practices (such as insufficient testing or over-reliance on automation without oversight in deployment), but we also note that those are also already covered under categories pertaining to cybersecurity (propensity to bypass safeguard) or risks of major accidents (misalignment with developer’s intent, e.g. through a misspecified objective function by the developer). Other examples, including e.g. self-replication or “intentionally” evading human oversight, seem to rely on unfalsifiable assumptions about the link between “capabilities” and negative outcomes of an anthropomorphized system. Keeping the category as is will at best require GPAI developers to go through an onerous compliance exercise with little to no benefit, and at worst increase the risk of more likely harms from GPAISR as a red herring. Either scenario creates a disproportionate barrier for small and open developers who rely on transparent, verifiable approaches and cannot engage meaningfully with ill-defined categories. Large developers may absorb the cost of bespoke internal processes to demonstrate partial alignment with these categories, but smaller actors, who are essential to innovation and research, cannot do so without clear standards and plausible evaluation pathways.

Overall, the taxonomy over-represents risks that reflect specific narratives put forward by a handful of large developers, without broader validation or alignment with Recital 110, which emphasizes risks to public safety, democratic processes, and critical sectors. Crucially, it leaves out systemic risks with well-documented harm mechanisms, such as biased decision-making in welfare systems or cascading infrastructure failures, that are

more likely to occur and more actionable through collaborative mitigation efforts.

The current framing also discourages identification of emerging risks by making non-listed risks appear optional ("where it can be reasonably foreseen"), and adding them only after those proposed by developers. This undermines proactive reporting and responsiveness to real-world harms.

Risk identification should instead be rooted in model-agnostic evidence, measurable capabilities, and collaborative research. The Code would be better served by removing the current taxonomy and instead supporting iterative, transparent processes for identifying systemic risks, grounded in cross-stakeholder input and external expertise.

Taxonomy of Systemic Risks, Appendix 1.2: Other types of risks for potential consideration in the selection of systemic risks

To what extent do you agree with this appendix?

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

Please explain your rating and suggest improvements

Many of the risks in appendix 1.2 are core to the systemic risks by AI, and align well with Recital 110. In particular, the following risks should be core to the systemic risk framework:

risks of major accidents and to critical sectors/infrastructure, risk to privacy, risk to non-discrimination, risk to environment

These risks should be listed in Appendix 1.1, making them mandatory for all deployers to test for.

Submitted by:
Lucie-Aimée Kaffee, EU Policy Lead & Applied Researcher, Hugging Face
Yacine Jernite, ML and Society Lead, Hugging Face