Hugging Face Comments on the First Draft of the GPAI Code of Practice

[First Draft of the General-Purpose AI Code of Practice published, written by independent experts | Shaping Europe's digital future]

Overall Code of Practice Draft Comments

While the Code of Practice Draft addresses some of the core challenges, there are a number of challenges yet to be addressed as well as problematic framing, making it hard to implement the code of practice or to base it on scientific evidence.

We find the Transparency measures to be the most mature part of the document, and provide feedback here on how to make them more actionable and inclusive of open research and development, we note that for those, as for testing requirements, the level of detail and guidance from the AI Office will make the difference between measures that work for the entire ecosystem and measures that fail all categories of stakeholders except for the largest incumbents. While we appreciate the approach of going from general to specific through subsequent drafts, we urge the writers to provide sufficient opportunity to discuss operational details, especially with a view to making them future-proof.

Conversely, the sections on risk taxonomies and assessment require the most work. The current taxonomy focuses on a narrow set of risks that are somewhat disconnected from both the actual likely risks of the technology and in some cases from scientific evidence. Risk taxonomies need to be significantly updated to address the full range of risks initially covered by the AI Act, ensure that development priorities reflect real-world impact, and give sufficient voice to stakeholders outside of the development chain; who have less of a vested interest in how the models are presented and advertised.

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

Transparency, Measure 1: Documentation for the AI Office

The current draft outlines important categories for the documentation. However, the items on data transparency and testing require significant further clarification to ensure sufficient opportunities for working group members to meaningfully contribute to the outcome. Additionally, the role of public documentation or public availability of the code supporting some

of the sections of the table, should be taken into further consideration. In cases where sufficient information is available, requirements of reporting to the AI Office should be lifted or alleviated

For data transparency requirements, we refer to the following proposal as an example of the categories of information that should be shared with the AI Office and preferably publicly disclosed:
https://openfuture.eu/blog/sufficiently-detailed-summary-v2-0-of-the-blueprint-for-gpai-training-data/

Please provide suggestions on how to improve this measure

For both Measure 1 and Measure 2, public disclosure should be encouraged to the extent possible, as it will greatly facilitate the adoption of common good practices. Barring that, the AI Office should provide relevant researchers easy access to the submitted documents. Given the fast pace of technical evolution, sufficient public access is necessary to allow relevant stakeholders and researchers across disciplines to ensure that documentation addresses their needs. Public disclosure is also significantly more accessible to developers of open models. Open release of sufficient documentation on platforms such as Hugging Face or Github should be explicitly described as satisfying the needs of disclosure to both the AI Office and to downstream model providers.

The section on data transparency is most in need of clarification in the proposed table. In particular, its relationship to the ”sufficiently detailed data summary” (Article 53 (1) (d), recital 107) needs to be clarified as soon as possible, to avoid unnecessary duplication of effort and support more meaningful documentation. We strongly encourage the working groups to provide a detailed proposal in the next draft to support sufficient working group involvement.

The qualification “where applicable” should be added to the documentation items about methods of distribution, interaction with external hardware/software, and technical means for integration. In models that leverage open-source software, such as the Hugging Face OSS libraries, that information is often better addressed in the existing software documentation.

The AUP focus on intended and acceptable use should be preserved in further drafts. Monitoring should be understood to be one tool among others, and not necessarily the most applicable or rights-respecting: developers should explain “**whether**, why, and how” they monitor user activity. If they do, the impact on user privacy should include details on the management and re-use of user data.

What KPI would you add for this measure?

Information disclosed to the AI Office should be evaluated based on how well it supports additional research and rule-making work by the AI Office and its partners who may access the data, including its scientific council and civil society representation. Any update mechanism should enable feedback from these stakeholders.

Transparency, Measure 2: Documentation for downstream providers

Comments on Measure 1 also apply to Measure 2.

Additionally, we would like to see greater clarity about why specific categories are meant to be disclosed to only the AI Office or only downstream providers. In particular, information about the training objective, testing conditions, and inference costs are of relevance to downstream providers, including for assessing fitness of purpose and reliability of systems.

Please provide suggestions on how to improve this measure

In the first draft, several categories of information that are often important and sometimes necessary for downstream providers to deploy systems in a safe and secure fashion are currently only noted as having to be disclosed to the AI Office.

In order to address fitness-for-purpose of AI systems, downstream providers need to assess several properties of the model, including how well its training data represents their activity domain, how closely the model's initial testing conditions match their use case, or whether data they might want to use to test systems for their own applications might have been included in the GPAI's training data at any stage. This information is particularly important for downstream providers whose products may be used in safety-critical domains, including settings that may not be considered high-risk applications under the Act (Article 6 (3)). Additionally, downstream providers who rely on a particular system need some understanding of the production cost of the system, especially in cases where those costs might be temporarily absorbed by the GPAI developers but subject to significant later increase.

Information disclosed to the AI Office about the sizes of different data sources, the training procedure and objectives, the energy consumption at inference time, and the exact testing procedures should also be made available to downstream providers to enable development that better protects the safety of people affected by the systems.

What KPI would you add for this measure?

In order to ensure that this Measure fulfills the information needs of downstream providers, we ensure establishing a mechanism for downstream providers to submit questions about the GPAI models they use in their commercial or research activity that is registered with the AI Office. Knowing which of these questions are covered by the categories and which are not would be invaluable in shaping the evolution of the transparency requirements.

For the items listed in the table, how should the Code of Practice provide greater detail?

For the model architecture and parameters, the draft should outline how providers of systems with API-only access should report the full stack involved, including when multiple models are used within an inference call, or how inputs or outputs to the model may be modified, especially in ways that might have bearing on the downstream provider's intent.

For the training, testing, and validation data, see our comment above. At a minimum, the documentation should additionally outline for what purposes various datasets are used, data processing steps aimed at improving model safety and security, and contamination analysis between testing and validation datasets and training datasets as available to the developer.

For the testing process, documentation should ensure that the findings are replicable and comparable across systems, including by providing the exact testing setup and a sufficient subset of the testing dataset to enable external reproduction and “apples-to-apples” comparison. To that end, the use of public benchmarks for a representative subset of performance and safety measures should be encouraged.

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

Copyright-related rules, Measure 3: Put in place copyright policy

Please explain your rating to this measure

The Measure as currently proposed presents a significant risk of excluding small and medium actors from participating in GPAI development, while its contribution to making larger actor’s practices more rights-respecting and fairer to creators of content under copyright remains uncertain.

Sub-Measure 3.1 requires developers to draw up a copyright policy, but does not provide meaningful information about what constitutes a valid such policy. Smaller teams, collaborations, and organizations that work on parts of the development chain in various ways in a distributed fashion, including by leveraging or publishing open models, tools, and datasets, will find this disproportionately difficult without stronger guidelines compared to the few best-resourced developers who own their entire development chain – as the former typically have access to fewer centralized legal resources, a greater diversity of questions to answer, and less ability to absorb the cost of fines if a good-faith effort is deemed insufficient *post hoc*. Given the role of open and collaborative research and development in supporting both innovation of a greater variety of techniques and investigation into existing technology, providing copyright holders with more clarity and options when defending the value of their content, we strongly encourage future versions of the draft to better address their requirements (and contributions).

Sub-Measures 3.2 and 3.3 also require significant further clarification to be manageable by smaller entities and developers working in open and collaborative settings. These submeasures should address in particular information flows between open datasets and developers, clarify whether datasets or models made available under a given license or with a simple signed user agreement constitute a contractual relation for those purposes.

Please provide suggestions on how to improve this measure

Sub-Measure 3.1 should provide significantly more detail on what constitutes a valid copyright policy, and enable actors to adopt such a policy by providing templates that address various development models. While we acknowledge that the development of the Code of Practice is

supposed to go from general to specific, we are concerned that waiting until the last draft for a fully detailed proposal will not provide sufficient opportunities for inputs from the most directly affected stakeholders. At the very least, the next draft needs to outline which parts of the “entire lifecycle” are deemed relevant and should be the focus of the policy, and possible policy items for both closed and open development and sharing.

Sub-Measure 3.2 should clarify whether the use of a publicly available dataset released under a content or software license or “click-through” user agreement is understood as a contract in this context. If that is the case, it should acknowledge that dataset developers may not be available for extensive interactions with each developer and provide guidance for interactions with these mostly static artifacts.

Sub-Measure 3.3 requires similar clarification of its scope. We do appreciate the focus on development choices and memorization measures and acknowledgement of the different dynamics for SMEs.

Copyright-related rules, Sub-Measure 3.1: Draw up and implement a copyright policy

While the idea of drawing up and implementing a copyright policy is valuable in principle, it presents significant challenges for SMEs and especially for open-source projects that often lack access to legal counsel. To address this, it is essential to provide a clear template and detailed guidance on what constitutes a satisfactory copyright policy. This should include how the policy's content aligns with the commitments outlined in subsequent sub-measures and what "implementation" entails, particularly over the long term. For open-source projects, which may no longer be actively maintained but remain valuable to the AI developer community, the guidance should consider realistic and practical solutions for sustaining compliance without imposing excessive burdens.

Please provide suggestions on how to improve this sub-measure

- Provide Templates and Guidelines: Develop a standardised template and detailed guidance for creating a copyright policy. This should clearly outline required elements, such as lifecycle compliance, and explain how the policy integrates with commitments in other sub-measures. Templates should be tailored to the needs of different stakeholders, including SMEs and open-source projects, to reduce the administrative burden.
- Simplify Requirements for Open Source Projects Introduce proportional obligations for open-source initiatives, recognizing their resource constraints and decentralised nature. For example, clarify that a policy could focus on the transparency of training datasets and adherence to known opt-out mechanisms rather than requiring ongoing active maintenance when a project is no longer actively developed.
- Facilitate Access to Expertise: Provide access to shared resources, such as legal guides or pro bono legal assistance programs, to help SMEs and open-source contributors navigate copyright compliance without the need for in-house legal teams.
- Specify Long-Term Maintenance Expectations: Include provisions for how copyright policies should apply to projects no longer actively maintained. For example, policies could include

instructions or disclaimers for users of the AI models, specifying how to address copyright concerns if no contact point is available.
- Ensure Alignment with Other Sub-Measures: Clearly distinguish how this sub-measure complements or differs from related requirements in the Code, such as transparency documentation or point-of-contact obligations, to avoid duplication or confusion.
- Establish a Sunset Clause for Compliance: For projects or models that are no longer actively maintained, introduce a sunset clause to limit the obligations of contributors, ensuring that the requirements are not indefinite for contributors who have ceased involvement.

What KPI would you add for this sub-measure?

- Number of Templates or Resources Accessed: Number of signatories who use the standardised copyright policy templates or other resources provided (e.g., legal guidelines, FAQs).
- Feedback and Improvement Requests: Number of feedback submissions or improvement requests received from developers (particularly SMEs and open-source contributors) regarding the copyright policy framework.
- Incident Reports of Copyright Infringement: Number of copyright-related complaints or issues reported that indicate non-compliance with the policy or inadequate implementation.

Copyright-related rules, Sub-Measure 3.2: Upstream copyright compliance

The code should clarify that this sub-measure applies only to contracts between data providers and users of datasets, excluding the reuse of freely licensed datasets. This is essential to align with the EU AI Act, which focuses on general-purpose AI models, not datasets. Extending the scope to freely licensed datasets would exceed the regulation's intent and undermine open licensing frameworks, which already address rights reservations. Datasets requiring a license agreement, such as the news corpora as in https://www.nytimes.com/2024/05/22/business/media/openai-news-corp-content-deal.html should be treated differently from freely available, openly licensed data. Users of freely licensed datasets would not have the necessary resources to do a due diligence on all freely licensed datasets, and also this could mean multiple streams of work on due diligence of the same dataset. Applying this measure to freely licensed datasets risks unnecessary burdens, discouraging open data use.

Copyright-related rules, Measure 5: Transparency

Copyright-related rules, Sub-Measure 5.3: Single point of contact and complaint handling

The sub-measure, as stated, lacks specificity as to the kinds of complaints that developers should prepare to handle. We note that existing complaint mechanisms like DMCA or GDPR-related requests work in great part because they have clear legal grounding,

definitions of what constitutes a valid complaint, and prescribed actions to follow up on the complaint. Without those, the designed point of contact will be much less efficient.

Additionally, for open source projects, many of the models are still available after the funding or organisational support for the project ended. Similar to sub-measure 3.1, it should be clarified here how to proceed when projects end or there is no clear organisational structure in the first place.

Please provide suggestions on how to improve this sub-measure

- Provide a template for a unified complaint format, including guidelines on what constitutes sufficient information to identify the validity of the claim, ownership of the work, or grounds for protection of the subject matter.
- Given a valid complaint, provide guidance on what constitutes “appropriate complaint handling procedure”, especially in terms of handling of both copies of the work held by the developers and models trained on datasets including copies of the work.
- Specify Long-Term Maintenance Expectations: Include provisions for how copyright policies should apply to projects no longer actively maintained. For example, policies could include instructions or disclaimers for users of the AI models, specifying how to address copyright concerns if no contact point is available.
- Establish a Sunset Clause for Compliance: For projects or models that are no longer actively maintained, introduce a sunset clause to limit the obligations of contributors, ensuring that the requirements are not indefinite for contributors who have ceased involvement.

Copyright-related rules, Sub-Measure 5.4: Documentation of data sources and authorisations

The relationship between this documentation and other data documentation requirements, including for Measures 1 and 2, should be further clarified. This sub-measure should clearly define how these data sources should be documented, to what level of detail they have to be provided, and further extend whether there are other relevant information beside copyright-related questions that should be documented and subsequently made available by AI providers.

Section III: [Working Group 2] Taxonomy of Systemic Risks

Taxonomy of systemic risks, Measure 6: Taxonomy

We are deeply concerned by the current focus of the risk taxonomy. The taxonomy in the first draft overrepresents a narrow subset of the risks described in the AI Act; a subset that focuses on the least defined and immediate risks without sufficient scientific grounding or precedent in broader technology. We urge the next draft to bring the focus back to

"reasonably foreseeable negative effects" such as "major accidents" (Recital 110) or irresponsible deployment in specific contexts.

Please provide suggestions on how to improve this measure

The narrow focus of the current taxonomy reflects disproportionate attention to intentional misuse and capabilities as naturally emerging model properties to measure post hoc. A more holistic approach should first recognise that recent examples of the systemic risks considered (e.g. CrowdStrike outage) have stemmed from immature deployment in safety-critical systems, while new "capabilities" have come from intentional design choices, primarily the inclusion of new kinds of training data.

Taxonomy of systemic risks, Sub-Measure 6.1: Types of systemic risks

Manipulation cannot be measured at the model level without a specific distribution model. CBRN risks are currently remote to inexistent, and requires grounded security research by domain experts, not highly abstracted evaluation by model developers. Automated use of models for research is part of their primary benign use and may at most be considered an accelerating factor for technology development, not a risk. Loss of control is primarily a risk tied to system design, not capabilities.

Please provide suggestions on how to improve this sub-measure

Cyber offence should be replaced by risks to cybersecurity, including through spread of code vulnerabilities. Categories mentioned above should be stricken. Categories of risks should cover a broader set of considerations from the AI Act, including consequences to public health, civic/human rights, and democratic processes stemming from model reliability fairness, as well as measurable risks to economic domains and communities (Rec. 110). See also: https://arxiv.org/abs/2306.05949.

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

Systemic risks included in the taxonomy should be risks that:

- Have an identified and minimally likely negative impact on a given social system (e.g. education, information ecosystems, journalism, labor, etc.)
- Are tied to specific development choices or properties that can be characterized using reproducible, falsifiable, and context-sensitive evaluations
- For scientific principles for risk assessment, see e.g.: https://ec.europa.eu/newsroom/dae/redirection/document/110171

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

Risks to the security or reliability of critical systems, including access to health, education, government benefits or essential services. Risks to democratic processes through the integrity of journalism or information ecosystems. Risks to the environment through increased dependence on energy-intensive technology.

How should the taxonomy of systemic risks address AI-generated child sexual abuse material and non-consensual intimate imagery?

AI-generated CSAM constitutes a systemic risk as defined insofar as it puts pressure on existing systems that are designed to protect targets and limit the spread of content. For specific harm mechanisms for CSAM harms, the importance of focusing on deployment and development choices, and the difficulty of anchoring work on "model capabilities", see e.g. https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf

Taxonomy of systemic risks, Sub-Measure 6.2: Nature of systemic risks

Origin, actors, and intent all require significant further elaboration.

Probability-severity ratio raises questions about who gets to define severity when different stakeholder groups are differentially affected and what constitutes a valid model to assess likelihood. It should also be understood that a risk is inherently defined by a notion of likelihood (a hazard by itself does not constitute a risk)

Please provide suggestions on how to improve this sub-measure

Origin should provide a definition of model capabilities. Model distribution should be inclusive of model development and deployment choices.

The actors driving the risks are currently missing the model developers, who are currently the ones in the strongest position to mitigate (or increase) risks through development choices. In particular, misrepresented or improperly documented models or capabilities are a strong risk factor.

Taxonomy of systemic risks, Sub-Measure 6.3: Sources of systemic risks

The sub-measure includes several notions that are not supported by evidence or that do not correspond to a quantifiable or falsifiable measure, including: self-replication, "power-seeking", "collusion", or a notion of alignment with universal "human values".

Please provide suggestions on how to improve this sub-measure

Several inclusions in the current draft are very relevant and should be further expanded, including notably considerations of number of business users and technology readiness, lack of reliability and security including biases, and confabulation in deployment settings with high precision requirements.

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

Measures/Sub-measures Specific Feedback on Section IV: [Working Group 2] Risk assessment for providers of General-Purpose AI Models with Systemic Risk

Risk assessment, Measure 8: Risk identification

The current version of the draft makes model developers responsible for identifying risks, selecting evaluations, and interpreting those evaluations. This raises two main concerns. While developers have the best understanding of some of the technical aspects of their models, they typically have neither the broader expertise (sociological, domain expertise, etc...) nor the legitimacy to decide which risks should be prioritized, especially when choices affect stakeholder groups differently.

Please provide suggestions on how to improve this sub-measure

Risk identification should occur outside of the development chain. For developers training models with systemic risks, an external review system, possibly with the AI Office scientific advisory board, should determine which risks should be evaluated based on parts of the information described in Measures 1 and 2 that can be disclosed ahead of deployment. We note that openly released models in particular provide additional opportunities for external stakeholders and experts to identify risks.