HUGGING FACE

Template for Submissions to the Multi-stakeholder Consultation FUTURE-PROOF AI ACT: TRUSTWORTHY GENERAL-PURPOSE AI

Authors: Yacine Jernite, Lucie-Aimée Kaffee

Working Group 1 - Transparency and copyright-related rules

The first Working Group focuses on detailing out documentation to downstream providers and the AI Office on the basis of Annexes XI and XII to the AI Act, policies to be put in place to comply with Union law on copyright and related rights, and making publicly available a summary about the training content. ***Please enter your comments below. This section is ideal for providing insights that may not be captured by the structured input above or provide any contextual information that may be relevant to this Working Group.***

Transparency requirements are particularly apt to contribute to more trustworthy GPAI while fostering innovation from a broad set of actors, especially in research institutions, start-ups, and SMEs.

The content of the data used during the training of a GPAI model data plays a substantial role in ensuring that the technology is robust, rights-respecting, and trustworthy. Information about training data provides necessary context for interpreting a model's results on performance benchmarks, identifying risks related to privacy, understanding the role and contribution of content under copyright in the technology, and making informed choices about deployment in specific contexts; as well as shaping which uses of a model, benign or malicious, will have better or worse performance. At the same time, templates for developers to provide this information need to contend with the scale and diversity of datasets involved. Commercial developers of GPAIs have also argued that excessive details on their full data pipelines could encroach on their trade secrets.

In order to manage this tension, transparency requirements should focus on the boundaries between the activities of the GPAI developers and external stakeholders. These would include:

- The sources of the data used in the various pre-training, fine-tuning, and evaluation datasets – including publicly available data, data obtained through licensing deals, and data elicited from crowd workers or users of the system. The initial conditions in which external data are acquired determine privacy and intellectual property risks, market dynamics around the value of creative works and data, and the ability of external stakeholders to broadly know when they might be entitled to exercise data rights.
- Extensive details on the evaluation datasets for publicly released performance results and other measures of model behaviors to qualify the scope, possible limitations, and

construct validity of the evaluations. Reproducible evaluations on broadly accessible benchmarks should be prioritised whenever possible.
- Details on risk mitigation strategies, especially through techniques such as training data filtering or content guidelines-based fine-tuning (sometimes called Constitutional AI) that raise many of the same questions as content moderation at scale, and thus require similar levels of transparency to enable proper governance as are requested of especially VLOPs in the European Union.
- Clarity on the uses of data obtained while running the model, such as documents used as inputs for a served model or user queries.

A focus on the above categories of information helps tailor transparency requirements to specific risks to EU citizens' rights and legitimate interest. It also preserves trade secrets that are more closely tied to the developer's exclusive handling of data. Notably, recent models have been reported to reach increasingly higher performances on benchmark thanks to more elaborate uses of "self-play" synthetic data and reinforcement learning, which would remain beyond the scope of the proposed transparency requirements.

Transparency-based approaches to GPAI governance present the unique advantage of being well aligned with the ethos of open-source and open science development, and the thriving ecosystem of start-ups and SMEs they support. Indeed, good documentation of proper uses and trust building through maintenance and transparency are instrumental to the success of OSS software. Free and Open Source AI developers and start-ups and SMEs also have unique constraints that should be acknowledged when formulating transparency requirements; by focusing on information the developers have access to and minimising the burden of engaging external certification or maintaining several versions of documentation, among others. For open models that are put on the EU market and shared on open repositories such as GitHub or Hugging Face, requirements should be manageable simply by publishing sufficient documentation alongside the model code or weights, so as to ensure that less-resourced organisations, or organisations that only exist for a limited times – e.g. punctual collaborations such as BigScience – can sustainably meet their demands.

GPAI developers, especially in academic institutions and SMEs, also commonly make use of publicly accessible datasets; since those are accessible and may be analysed by external stakeholders, they should be understood to comply with transparency requirements by default, as long as the developers disclose their use.

Working Group 2 - Risk identification and assessment measures for systemic risks

The Code of Practice should help to establish a risk taxonomy of the type and nature of the systemic risks at Union level, including their sources. The second Working Group will focus on

detailing the risk taxonomy based on a proposal by the AI Office and identifying and detailing relevant technical risk assessment measures, including model evaluation and adversarial testing. Please enter your comments below. This section is ideal for providing insights that may not be captured by the structured input above or provide any contextual information that may be relevant to this Working Group.

Risk identification and assessment measures are an essential part of enabling foresight and mitigating risks of harms. In order to play that role, evaluations need to be properly scoped, scientifically validated, reproducible, and accessible to all categories of actors, including developers of open models, start-ups, and SMEs developing GPAI models and systems. In particular, the risks of GPAI systems should be defined primarily by the people most likely to be affected with sufficient involvement of external stakeholders to ensure construct validity.

In order to improve the state of the art in and adoption of risk assessment practices while avoiding fragmentation, the Code of Practice should focus on having developers:

- Participate in public efforts involving civil society and public organisations in establishing consensus-driven and scientifically validated risk evaluations, with a view to contributing to international standards. Risks depend on the context of an AI model and its users, therefore it is crucial to include a variety of perspectives in their definition [1].
- Contribute to the development of public benchmarks and risk evaluation methodology according to their size and capacity.
- Collaborate on joint infrastructure for running those evaluations, leveraging open source software and open models as appropriate to facilitate broad participation.

Models that are released on a Free and Open Source (FOS) basis, in particular, should benefit from particular support in meeting these requirements. Requirements for FOS GPAI models should prioritise evaluations that can be run without significant extra computation budgets or involvement of third-party organisations to acknowledge both their default increased transparency (more expensive or involved evaluations may be run by external parties as needed pre-deployment in specific commercial products) and different operational constraints (organisational dynamics for academic actors or collaborations between open developers may not allow for costly external audits).

[1] Wachter: Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond
https://ora.ox.ac.uk/objects/uuid:0525099f-88c6-4690-abfa-741a8c057e00/files/sht24wm314

Working Group 3 - Risk mitigation measures for systemic risks

The Code of Practice should be focused on specific risk assessment and mitigation measures. The third Working Group will focus on Identifying and detailing relevant technical risk mitigation

measures, including cybersecurity protection for the general-purpose AI model and the physical infrastructure of the model. Please enter your comments below. This section is ideal for providing insights that may not be captured by the structured input above or provide any contextual information that may be relevant to this Working Group.

Risk mitigation measures for systemic risks should focus on the entire development process, from project design to data curation to final fine-tuning and deployment. While measures based primarily on fine-tuning and deployment guardrails, such as input or output filtering, can be a part of lowering overall risks of especially unintentional misuses, they have been shown to be too brittle to constitute a complete solution, subject to e.g. jailbreaking and unintentional removal through fine-tuning (including through closed APIs) [1,2].

Mitigation strategies that focus upstream on the development process, including through training data curation or early-development considerations on the trade-offs inherent in different capabilities, on the other hand, present the dual advantage of being more robust and more accessible to well-intentioned actors training models to be released openly.

In order to facilitate the development of more robust systemic risk mitigation strategies while preserving the ability of responsible developers to share open models that support research and innovations, we recommend that the Code of Practice focus on:

- Prioritising safety by design approaches along the full development chain
- Providing clarity on which risk mitigation approaches should be leveraged by developers or by deployers
- Contributing to open-source tools for risk mitigation that can easily be adopted by actors of all sizes, and can be externally validated, especially where they require trade-offs between different values
- Facilitate skill-sharing and constitution of a set of best practices by encouraging transparent reporting on safety strategies - risk mitigation is in the public interest and should not be treated as an exclusive competitive advantage, sharing information between actors helps spread good practices and identify unforeseen effects faster.

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

General Considerations for the drawing-up of the Code of Practice

Please provide any general considerations that should be taken into account during the drawing-up of the first Code of Practice for providers of general-purpose AI models.

The Code of Practice represents a unique opportunity to catalyse progress toward more trustworthy GPAI technology across actors. In order to meet those goals while realising the aims outlined in the AI Act of supporting research and innovation, including through free and open source AI, its design process needs to fully take the needs and strengths of all actors into consideration. This includes the most well-resourced developers deploying GPAI systems at scale, start-ups and SMEs who increasingly have access to the resources needed to train their own versions of GPAI that may be better suited to their own use cases (including different languages, domains, modalities), and non-profit and academic researchers who develop open GPAI models that enable much of the research. This research is needed to better understand those systems and ensure that regulation keeps pace with the latest technical developments – including e.g. the BigScience [1] and BigCode [2] efforts, AI2 work on the DolMA dataset [3] and OLMO language models [4], and EleutherAI's Pythia models [5].

In order to ensure that the Code of Practice prioritises measures that are both accessible to all of the above stakeholders and foster more robustness and accountability, we recommend that all working groups:

- Prioritise open collaboration with external experts to ensure that measures are driven by needs expressed across stakeholder groups
- Prioritise documentation and transparency as significant contributors to robust and reliable technology
- Focus on measures upstream in the development chain supported by open tools and methods
- Ensure that all measures are properly documented, and that evaluations in particular are reproducible and subject to external scrutiny
- Ensure that start-up and SMEs have a path to compliance that does not depend on extensive third-party certification or audits
- Ensure that requirements for developers of open models are suited to the part of the development chain they have direct control over

Such an approach will not only support innovation and diverse participation in the evolution of GPAI technology in the EU, but also ensure that the Code of Practice remains relevant as technical conditions evolve over the coming years.

[1] https://bigscience.huggingface.co/
[2] https://www.bigcode-project.org/
[3] https://allenai.github.io/dolma/
[4] https://allenai.org/olmo

[5] https://huggingface.co/EleutherAI/pythia-6.9b