Key opportunities for improvement in the EU AI Act's second draft of the Code of Practice for GPAI model providers

While the second draft Code of Practice ("Code") contains some welcome improvements and clarifications, substantial revisions are necessary to bring the Code into alignment with the safety and innovation-friendly framework that the AI Act has as its goal. Future drafts of the Code should more closely align with the scope of the AI Act, avoiding significant expansions or divergences from the legal text, and define more outcome-oriented and less prescriptive Measures. With substantial revisions, the Code could provide the clarity and streamlined approach needed to support effective compliance with the AI Act for an expanding European AI ecosystem, while ensuring transparent and responsible development and deployment of GPAI models.

The current draft transparency and copyright Commitments applicable to all GPAI models go beyond the AI Act's scope and include Measures that would undermine innovation without adding clear safety or other regulatory value.

- Overly detailed **technical documentation** expectations raise significant concerns around confidentiality, trade secret protection, information hazards, and impact on innovation. We recommend removing and aligning with the AI Act's legal text, disclosure of information that could:
  - Constitute **trade secrets** (e.g., description of how the model architecture departs from standard practices; information about "each step or stage" of model training);
  - Pose risks of misinterpretation given a **lack of standardized methods** for tracking and reporting (e.g., computational resources for model *inference*; detailed information about acquisition or processing methods to address risks of concern in training, testing and validation data); and
  - Risk **confusing downstream AI system providers**, therefore disrupting innovation (e.g., a list of allowed types of high-risk systems or "restricted tasks" at the model layer).

- Copyright provisions should adhere more strictly to the Act and be refocused on Measures that are designed to support model providers in **putting in place appropriate policies and procedures for, rather than proactively proving, compliance**.
  - Amend the Preamble to focus on clear obligations as set out in the articles of the AI Act, including, for example, the provisions in Article 53 1(c). Remove requirements based on **interpretations of the legal text**, that are not clearly grounded in legislative provisions and conflate recitals of the AI Act with enforceable obligations in the articles of the Act. Such requirements risk creating ambiguity and overextending the scope of Signatories' responsibilities under the Code.
  - Remove Measures 2.3 and 2.4 to avoid **requiring model providers to proactively demonstrate proof of compliance** and **reversing the burden of proof**, which would impose a significant administrative burden and potentially stifle innovation.

A significant portion of the draft Commitments applicable to GPAI models with systemic risk are overly prescriptive, locking in rapidly evolving safety and security practices, or duplicative, resulting in redundant expectations that risk diverting safety resources. The Code should be reviewed with the goal of ensuring that all Measures and proposed KPIs are as streamlined and outcome oriented as possible, strengthening safety and innovation outcomes by enabling appropriate flexibility and prioritization.

- Overly prescriptive expectations for model providers to commit to a specific **percentage of engineering hours** risk locking in current approaches to risk assessment rather than incentivizing the development of rigorous new techniques and automation tools for challenges of scale, such as red teaming. More outcome-oriented approaches would also allow model providers to select the most suitable methods to advance safety depending on the model and internal context.

- Overly **prescriptive expectations for security mitigations**, including how model providers would need to protect stored model weights and related assets, should be replaced with more risk-based and outcome-oriented security measures, including to reflect distinctions between the risk profile of covered assets. Where feasible, deferring to relevant technical standards[1] for further implementation detail will support a more flexible, risk-based, and future-proof approach.

- Expectations to report on adherence and adequacy assessments every six months are not risk-based and will divert AI safety engineering and research resources towards producing **redundant documentation**. They should be removed in favor of risk-based commitments to report on updates to Safety and Security Frameworks and to provide updates to Safety and Security Model Reports, as deemed necessary based on major shifts in a model's risk profile.
- Expectations to provide, upon request, external evaluators with sufficient time, engineering support, compute budgets, and "non-restrictive" access to models go far beyond the still-evolving norms of providing select assessors with access to models pre- and/or post-deployment under strict guidelines and with security and confidentiality restrictions in place. The Code should instead **recognize a variety of approaches to leveraging external assessments post-deployment**, including via approaches that are **not dependent on non-restrictive access**, such as external research, responsible disclosure, and bug bounty programs.

Several draft Commitments go beyond the AI Act's and thereby the Code's intended scope of GPAI models by implicating AI systems. The Code should clearly scope its expectations for systemic risk evaluation and mitigation to practices that can be implemented exclusively at the model layer, rather than expanding commitments to cover the downstream systems layer. This will improve clarity and appropriately apply value chain responsibilities, supporting innovation and achievement of desired regulatory outcomes.

- Remove provisions for model providers to **evaluate a model's capabilities and limitations** for all existing and future system-level deployment scenarios "relevant" to a risk being assessed. The Code should instead explicitly set expectations on how model-level risk assessment and mitigation, must work in concert with system-level risk assessment and mitigation, already covered by the AI Act's requirements for high-risk systems.
- Limit provisions on **model-level post-deployment monitoring** that directly implicate the system level and can also present significant privacy challenges, as well as conflict with requirements for highly regulated global customers. Receiving and investigating reports from system providers and deployers and actioning those reports as appropriate would suffice.

[1] For example: ISO 27001, ISO 27017, ISO 29147, ISO 30111, Trusted Computing Group (TCG) standards, NIST SP 800-53, NIST SP 800-171, and INCITS 359-2004.