Key Considerations on the third draft of the GPAI Code of Practice

While we welcome several improvements in the third draft, including stronger alignment of the Transparency Section with the AI Act, it still contains provisions that are overly complex, inflexible, beyond AI Act requirements, or inconsistent with ecosystem readiness for effective implementation. Such provisions would draw upon finite risk management resources without having clear added value for or necessarily being consistent with desired regulatory outcomes.

We continue to see the potential for the Code to provide an agile and practical method for model providers to demonstrate compliance with AI Act requirements. Targeted but critical improvements, along with further streamlining and simplification, will however still be needed to ensure the Code is not only firmly grounded in the legal text and advances regulatory outcomes, but also constitutes a practical compliance mechanism. To that end, our key recommendations are summarized below in Section I, followed by a more detailed description of those concerns and suggested amendments in Section II.

I. Summary of Key Recommendations

Measures for all GPAI Models

- Remove or revise documentation form obligations that go beyond AI Act requirements and implicate trade secrets without clear value for regulatory outcomes, e.g., to document the details of model architecture innovations; data points for training, testing, and validation data in exact numbers versus ranges; methods and models for synthetic data generation; and compute consumption in the inference stage.
- Revise the Code's requirements for identifying and complying with rights reservations when crawling the World Wide Web by removing limitations on how the opt-out can be expressed to the robots.txt exclusion protocol and other standards that are developed.

Measures for GPAI Models with Systemic Risk

- Streamline and simplify the Code to support risk-prioritized compliance.
- Align external assessment Measures with the AI Act, which allows for either internal or external testing, by clarifying their voluntary nature, and take a less prescriptive approach to collaboration with external experts.
- Remove overly prescriptive internal assessment requirements, including model-specific adequacy assessments as they overlap with other, more effective Measures, and overly detailed credentials of personnel that make up internal "model evaluation teams", in recognition of the value of diverse expertise for devising and running evaluations.
- Focus all Measures on the model level by avoiding conflation with system-level risk evaluation and management, which the AI Act regulates as AI system obligations, and by limiting the "systemic risk taxonomy" to risks that can be assessed at the model level.
- Limit multilingual evaluation expectations to languages providers claim to support.
- Remove reporting on "near-misses", limiting the scope to confirmed serious incidents.

II. Context and Suggested Amendments

Measures for Providers of all GPAI Models

Within the model documentation form, we recommend revising or removing several categories that go beyond the AI Act and/or risk disclosure of trade secrets without clear value towards regulatory outcomes, including: (1) how a GPAISR's architecture deviates from standard model architecture; (2) the number of data points in two significant figures rather than ranges; and (3) a detailed description of methods and models used for synthetic data generation.

- Suggested amendments:
  - **Replace** "*where model architecture departs from standard architecture" with "key impacts to model behavior or capabilities that stem from architectural choices"*.
  - **Allow** providers to indicate *ranges versus exact numbers* for the reporting of unit and number of data points for training, testing, and validation data.
  - **Remove** obligations on methods and models used for synthetic data generation.

The requirement to document a model's compute consumption in the inference stage goes beyond the AI Act, which only concerns compute used in model development.

- Suggested amendment: **Remove** the obligation to report benchmarked amount of compute used for inference costs.

Obligations to describe methods in data acquisition or processing to address harmful content and prevalence of personal data in training data are inconsistent with the AI Act, which requires description of methods to detect "unsuitability" of data sources.

- Suggested amendment: **Narrow** requirements for describing data acquisition or processing methods to those for which there is technical feasibility and unsuitability can be determined with more legal clarity at the model level (e.g., CSAM).

Measures for Providers of GPAI Models with Systemic Risk

We recommend that the Code is streamlined and simplified, supporting greater clarity and more prioritized compliance. Commitments and Measures impacting GPAISR providers remain lengthy, overlapping, and convoluted. For example, Commitments II.2, II.4, II.6, II.7, and II.11 focus on how providers should assess and mitigate risks, with Commitments II.1, II.5, II.8, and II. 9 also detailing related expectations for when and how providers should conduct assessments, including of the effectiveness of mitigations, to make deployment decisions.

- Suggested amendment: **Streamline** and simplify the text to the greatest extent feasible, integrating overlapping Commitments (and Measures), such as II.2, II.4, II.6, and II.7.

We recommend that the Code encourages, rather than requires, GPAISR providers to leverage external assessment. While the AI Act requires model providers to evaluate and assess risks, it does not mandate a specific method for doing so, and Recital 114 explicitly allows providers to choose internal *or* external testing as appropriate. Moreover, AI evaluation science and standards are nascent, expert external assessors of model-level systemic risks are few, and substantiated criteria to have assurance of the validity of external

tests is still being developed. Mandating external assessment thus requires use of limited risk management resources without clear value towards regulatory outcomes, creates legal uncertainty that could impact innovation, and goes beyond the scope of the AI Act.

- Suggested amendments:
    - **Align** Measures II.11.1 and II.11.2 on external assessment before and after market placement with the AI Act by clarifying that external assessment is an optional risk assessment method rather than a requirement. External assessment before market placement should be encouraged when the following conditions are met: the GPAISR in question poses additional risks compared to existing GPAISR; the GPAISR provider lacks sufficient internal expertise to assess relevant systemic risks; and the GPAISR provider can find qualified external assessors.
    - **Encourage** flexibility in methods for external assessment or collaboration with external experts on risk assessment and mitigation in Measure II.4.11, replacing an emphasis on white- or grey-box access to models or for access to specific data, compute, or assessment resources. A more flexible approach could also encourage external contributions to tests run internally, policies and practices for managing responsible disclosure, and use of bug bounties. Likewise, adjust the framing of any qualifications for external evaluators in Measure II.4.11, recognizing that standards have not yet been developed to assess their readiness and quality, and that defining norms before there exists more meaningful evidence of directly relevant criteria involves risks, such as reducing openness to a diversity of assessor profiles.

We recommend removal of the requirement for model-specific adequacy assessments due to their clear overlap with more effective risk assessment measures during model development. Model-specific adequacy assessments would result in a premature, incomplete, and misleading picture of model reliability, diverting resources with limited potential value towards regulatory outcomes. Specifically, Measure II.9.2 would require model providers to submit detailed information about capability forecasts, potential systemic risks, assessment and mitigation plans, and systemic risks that may arise during the development phase—just four weeks after the initial GPAISR notification. In most cases, this would mean that providers would need to submit this information before model training is finalized—and before post-training and evaluations, which enable a much clearer and more useful look into a model's risk profile, are complete. Multiple other Commitments require Signatories to a) implement their Safety and Security Frameworks throughout the model development process, which should suffice as a commitment that appropriate mitigations will be implemented if a significant risk materializes during pre-training; and b) provide risk assessment and mitigation information via Safety and Security Model Reports that Signatories will submit to the AI Office upon placing a GPAISR on the market.

- Suggested amendment: **Remove** Measure II.9.2 and integrate appropriate portions into Safety and Security Framework and Safety and Security Model Report Commitments.

We recommend removal of prescriptive requirements related to qualifications for internal "model evaluation teams." As in the context of external assessment, proposed qualifications (e.g., PhD degree) reflect a particular assessor profile without being grounded in standards for measuring assessor readiness or other evidence of impact. In addition, such requirements fail to acknowledge that many teams with diverse profiles are involved in

devising and running evaluations across large organizations; for example, research teams may devise evaluations while security or other engineering teams may run them.

- Suggested amendment: **Remove** prescriptive requirements for qualifications of internal evaluation teams under Measure II.4.11 as criteria should remain optional until verifiable standards are developed, allowing for more openness to a diversity of assessor profiles.

We recommend that the Code avoids conflating model- vs. system-level risk assessment and removes the requirement for GPAISR providers to cover system-level deployment scenarios that are already regulated under the AI Act's AI system obligations. While model providers can offer tools and best practices to help meet the AI Act's requirements for high-risk AI systems (consistent with guidance that the AI Act specifies is to be developed outside of the GPAI Code of Practice), the Code should not obligate model providers to conduct, mandate, or report on system-level evaluations.

- Suggested amendment: **Remove** Measure II.11.1, obligating model-level assessments focused on general risks and capabilities, with system-level context viewed only as an optional risk management indicator.

We recommend removal of harmful manipulation from Appendix 1.1, **limiting the systemic risk taxonomy to risks that can be assessed at the model level**. Selected systemic risks in the taxonomy need to be specific, clear, and assessable at model level. Broadly understood, harmful manipulation is not specific to a model's high-impact capabilities and is contextual and/or heavily influenced by system-level deployment decisions—and therefore especially difficult to measure at the model level, as manipulative behavior is more typically associated with functionality and usability enhancements that emerge once a model is integrated into a system. In addition, this risk is already addressed through the relevant system-level prohibition on subliminal, manipulative or deceptive techniques under Article 5 of the AI Act.

- Suggested amendment: **Shift** 'harmful manipulation' to Appendix 1.2 or 1.3, as a risk factor for which consideration would be encouraged.

We recommend the Code provides flexibility regarding specific techniques for post-market monitoring, especially those that directly involve system-level risk management practices that overlap with broader AI Act obligations. Specifically, in addition to encouraging consideration of multiple potentially appropriate techniques, Measure II.4.14 obligates providers to monitor system-level deployment, either to detect breaches of use restrictions or as part of first-party systems, overlapping with the AI Act's prohibited practices and broader AI system obligations.

- Suggested amendment: **Encourage** consideration of a range of potential post-market monitoring techniques under Measure II.4.14, avoiding mandates of techniques that overlap with the scope of broader AI Act provisions.

We recommend limiting any expectations for multilingual evaluations of systemic risks to languages that the model provider claims to support. Availability of linguistic and culturally specific data is a known, cross-industry limitation. Model providers that aim to support non-English languages currently face challenges in performing multilingual evaluations due to limitations in available instruments to measure and mitigate systemic risks in non-English languages at scale. Given limitations, providers should focus resources on evaluations in languages they claim to support versus be expected to cover all major

European languages in evaluations of multilingual models, and development of multilingual evaluations should be coordinated at an industry-wide level.

- Suggested amendment: **Limit** scope of multilingual evaluations required under Measure II.4.8 to languages the GPAISR provider claims to support**.**

We recommend removal of obligations to keep track of, document, and report "near-misses," as this involves an unclear scope that goes beyond the AI Act, which only scopes in confirmed incidents. A focus on confirmed incidents supports greater legal clarity as well as risk-based prioritization of finite resources, including for incident response.

- Suggested amendment: **Make voluntary** any reporting that goes beyond confirmed serious incidents under Measure II.12.2**.**