EU AI Act: Key opportunities for improvement in the first draft of the Code of Practice for GPAI model providers

The implementation phase of the EU AI Act and corresponding secondary legislation present an important opportunity to provide the clarity needed for an expanding European AI ecosystem and the safety and innovation-friendly framework the Act has as its goal. The Code of Practice for GPAI model providers ("Code") exemplifies this opportunity in the immediate term. Microsoft describes below key substantive concerns regarding the initial draft version of the Code as well as recommended ways to address them. We recognize the significant work already invested in rapidly developing a high-quality draft, and we appreciate opportunities to contribute to the Code's improvement.

The Code should align with the letter of the AI Act, which was designed in accordance with the EU's goals of advancing safety and fostering innovation, by ensuring that future versions address the initial draft's significant divergences from the legal text. For example, the draft Code sets transparency expectations that go beyond the Act, implicating trade secrets without articulating additional safety value; puts forward a need for pre-deployment testing by the AI Office and independent third parties, despite a lack of ecosystem readiness to deliver consistently valid evaluations; and requires reporting of "near misses" (versus serious confirmed incidents).

The first draft of the Code introduces requirements for **technical documentation and information-sharing** that go beyond the scope of the AI Act and raise concerns around confidentiality, trade secret protection, and information hazards.

- Measures 1 and 2 on technical documentation go beyond the letter of the Act and include concerning expectations on the types of information that providers will need to make available to the AI Office (upon request) and to downstream providers, such as the proportions of data sources used for training, testing, and validation, and specifics on model architecture (e.g., number and types of layers), which go beyond the scope of Annexes XI and XII. Expectations to document both computational resources used for inference and energy use also go beyond the AI Act's Annex XI, which focuses on computational resources used for model training and allows for reporting estimated model-level energy consumption; such requirements also pose challenges due to the lack of standards for tracking and reporting energy use.
- Sub-Measure 13.6 implies an expectation for model providers to share detailed information about safety and security testing that could compromise the value of such tests for future models and risk assessments.

Recommendation: The Code should align with Annex XI and XII's defined scope in the AI Act legal text. Any additional information categories under the Code should be framed as optional implementation approaches rather than mandatory requirements. The Code should also provide clear context on the regulatory objectives for requesting such information so that model providers can propose alternatives that may be less sensitive but still responsive to the regulatory rationale behind requesting those categories of information. Sub-Measure 13.6 should explicitly acknowledge that the information provided in the "Safety and Security Report" (SSR) will not allow for independent assessment of any results, evidence, or analysis, but rather only enable assessment of the methodology itself, i.e., the risk assessment and mitigation process outlined in the "Safety and Security Framework" (SSF) and implemented in the SSR.

The draft Code would also effectively establish **a pre-market authorization regime**, exceeding the scope of the AI Act and contrasting with its emphasis on post-market monitoring and risk management at the model level:

- Sub-Measure 17.1 calls for ensuring sufficient independent expert testing before model deployment, e.g., by the AI Office and third-party evaluators.
- Sub-Measure 14.3 would similarly require signatories to detail in their SSF when development and deployment decisions will have input or require external authorization from external actors, including relevant regulators such as the AI Office.
- Sub-Measure 10.3 further implies mandatory third-party validation of all evaluation results, contradicting AI Act recital 114, which allows providers to conduct evaluations internally or externally as appropriate.

Recommendation: The Code should explicitly enable model providers the flexibility to perform pre-deployment evaluations with high scientific rigor and to validate evaluation results with internal and/or third-party experts, in line with recital 114. While Art. 92 empowers the AI Office to appoint independent experts to carry out evaluations on its behalf, this is limited to evaluations carried out in the context of investigatory actions. Clarifying this flexibility would ensure the Code does not go beyond the scope of the AI Act nor overly burden the EU market without clear risk management value, especially considering methods and processes for confirming the quality of third-party evaluation services will take time to put in place.

The draft Code, in the context of **serious incident reporting** in Sub-Measure 18.1, implies a requirement to also report near-misses. This goes beyond the scope of the AI Act, where reporting is only required for (confirmed) serious incidents, in line with existing practices under EU cybersecurity legislation, which is important for providing clarity and focusing risk management resources. In addition, the AI Act's legal text only defines serious incidents at the system level; the Code should identify criteria for model-level serious incidents.

Recommendation: The Code should remove reference to "near misses" and clarify that the scope of reporting is limited to "confirmed serious incidents", bringing it in line with the AI Act's requirement to report serious incidents.

The Code should distinguish between relevant systemic risks, and appropriate mitigations, at the model versus system level.

Sub-Measure 6.1 identifies **categories of systemic risk**, such as "persuasion and manipulation" and "large-scale discrimination", which are contextual and/or heavily influenced by system-level deployment decisions, and therefore especially difficult to measure at the model layer. Sub-Measure 6.3.3 lists several socio-technical factors beyond model capabilities and propensities, such as the potential for downstream users to remove guardrails, that are more typically associated with functionality and usability enhancements that emerge once a model is integrated into a system and are difficult to evaluate at the model level.

Recommendation: The Code should explicitly set expectations regarding how assessment and mitigation of systemic risks must work in concert with system-level risk assessment and mitigation, already covered by the AI Act's requirements for high-risk systems. Model providers should be given flexibility in determining what factors to consider and address as nature (Sub-Measure 6.2) and sources (Sub-Measure 6.3) of systemic risk.

Sub-Measure 11.4 on **post-deployment monitoring**, as currently conceived, would implicate platform- and/or system-level capabilities like monitoring production metrics (e.g., how often platform-level classifiers are triggered) or system-level capabilities like identifying where system outputs are not aligned with intended behavior—and thus assumes that model providers are also platform and/or system providers or that platform and/or system providers or system deployers report to model providers. Such monitoring also presents privacy challenges and conflicts with requirements for highly regulated global customers.

Recommendation: The Code should limit post-deployment monitoring requirements at the model level to receiving and investigating reports from system providers and deployers and actioning those reports as appropriate.

Sub-Measure 10.5 would require **evaluation of a model's capabilities and limitations** for all existing and future deployment scenarios. This requirement fails to acknowledge that model and system evaluations differ. Mandating system-level evals at the model layer imposes an unreasonable burden on model providers to anticipate and evaluate use cases they do not build for, fully have insight on, or have appropriate data or tools to assess.

Recommendation: The Code should remove this sub-measure, recognizing that the AI Act already addresses system-level risks through the application of the AI Act's high-risk AI system provisions. Model providers can support these efforts by providing tools and best practices, as outlined in Sub-Measure 10.8, but should not be responsible for conducting or reporting on system-level evaluations.

The Code should be reviewed end-to-end to make sure that all Measures and Sub-Measures are necessary, consistent, and add value towards achieving regulatory outcomes.

Measures **overlap or interconnect**, but these relationships are inconsistently acknowledged. For example, Measure 12 appears to rely on the "intolerable level" defined by model providers in Sub-Measure 9.3, though this link is not explicitly stated, and the connection between Sub-Measure 13.7 and Measure 14 is unaddressed.

Recommendation: The Code should draw out these and other intersections, which may also surface opportunities for streamlining and delivering greater clarity.

Measures 6-22 also **apply unevenly** across model and/or deployment scenarios.

Recommendation: The Code should allow model providers the flexibility to apply measures as appropriate to assess and mitigate identified risks. This flexibility should be clarified with explicit language in the Code's sections on governance.

The Code should reinforce the AI Act's risk-based approach by applying systemic risk requirements to the most advanced models demonstrating significant risks.

Regulatory efforts should be directed toward **mitigating significant risks**, applying to providers based on the risks posed by their models rather than on their size, consistent with the Act's risk-based approach. While onerous requirements applied to a wide range of models could affect what is made available on the EU market, today's most powerful models have been evaluated for dangerous misuse capabilities, and the findings from those evaluations have *not* identified unacceptable risks (e.g., see pre-deployment testing of Anthropic's upgraded Claude Sonnet 3.5 model, jointly conducted by the U.S. and the UK AI Safety Institutes). Rather than

establishing a framework that exempts model providers based solely on their size, regardless of the risks associated with their models, a risk-based approach would consider risks of today's most powerful models and focus in on a scope that addresses significant risks of concern. A focus on models that pose significant risks would also help alleviate concerns voiced by European start-ups and SMEs about the burdensome nature of the AI Act[1] while maintaining a principled risk-based and safety-oriented approach.

Recommendation: The Code should prioritize the application of risk assessment, testing, reporting, and notification measures for the most advanced GPAI models with systemic risk, defined as models that are trained with compute power over 10^{26} FLOPs and that demonstrate leading indicators of high-impact capabilities.

The AI Office should clarify the scope of the application of the Code to bring clarity to downstream entities.

The scope of application of the Code will depend on further guidance on what constitutes **substantial fine-tuning**. Clarifying this term will be crucial for downstream providers to understand whether and when they could be considered GPAI model providers and thus subject to the Code. Depending on this definition, additional entities may be brought into the scope of the Code without the opportunity to properly contribute to the drafting process. Further clarity will also be crucial to ensure legal certainty for companies across the AI value chain.

Recommendation: The AI Office should provide a further detailed definition for fine-tuning and/or thresholds for substantial fine-tuning as soon as possible through guidelines, welcoming input from model providers and deployers.

[1] In a recent survey by VC firm Atomico, over half of participating Europe-based tech start-up founders and investors responded that the AI Act has negatively impacted the conditions for starting or scaling up a tech company in Europe (p. 96).