Third Draft of the GPAI Code of Practice Misses the Mark on Fundamental Rights

18 March 2025

The AI Act envisages a strong role for fundamental rights in the Code of Practice

Risks to fundamental rights are a core aspect of the Code of Practice under the AI Act, with the Act's systemic risk definition expressly encompassing GPAI models with actual or reasonably foreseeable negative effects on public health, safety, public security and fundamental rights. The Act exemplifies some of these risks, explicitly outlining risks of harmful bias and discrimination with risks to individuals, communities or societies, as well as harming privacy (Recital 110). However, the latest version of the draft does not include any of these risks in the selected risk taxonomy, instead placing them in the optional risks list.

The latest version of the Code appears to be the result of a narrow interpretation of the AI Act – one which requires a cause-and-effect relationship between high-impact capabilities – a concept defined in the Act as capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI model (Article 3(64)) –and systemic risks, such that risks not directly resulting from those capabilities are excluded from the "selected" risks taxonomy. In practical terms, this means that fundamental rights risks which stem from GPAI models with systemic risks that do not arise from these models' comparative advantages in terms of sophistication or development will altogether be excluded from consideration in the draft Code, drastically reducing the avenues enabling scrutiny of these risks.

This de minimis approach is undesirable because it creates an additional layer of conditionality for systemic risks to be assessed – it is not enough for a GPAI model to cross into the systemic risk technical threshold for risks to be considered; it must also be established that those risks specifically relate to the high-impact capabilities of those models. This two-stage approach circumvents a robust and effective assessment of GPAI model risks, particularly as the mere fact that GPAI model meets the systemic risk thresholds set out in the Act makes it likely that, at least in an abstract sense, there is increased potential for high impact. The AI Act agrees with this approach, noting that thresholds should be "strong predictors of generality, its capabilities and associated systemic risk of general-purpose AI models" (Recital 111). Other aspects of the AI Act show a clear intent for GPAI model governance – and by extension the Code – to robustly cover fundamental rights risks. For example, a key obligation that the Code is intended to elaborate on is the obligation for GPAI model providers to document and flag serious incidents, which encompass infringements of EU law addressing fundamental rights (Articles 3(49), 55(1)(c)). Going even further, the Act empowers the European Commission to circumvent standards – including

those applying to GPAI models – and elaborate their own rules if they do not consider fundamental rights concerns to be sufficiently addressed by those standards (Article 41(1)). It is perplexing that failure to consider fundamental rights risks would constitute valid grounds for the Commission to reject standards, but that a similar failure by the Code would be acceptable.

The identification of systemic risks should not be strictly tied to the novel capabilities of GPAI models, but should instead consider the scale of impact from the most sophisticated, far-reaching foundational GPAI models. This view is supported by the AI Act, which states that systemic risks are understood to increase both with capabilities and reach. Because GPAI models meeting the threshold set by the AI Act are likely to be used as a foundation for a wide range of AI applications, the Code should approach all risks in the taxonomy, and in particular those to fundamental rights, taking into account the fact that GPAI models with systemic risk will exponentially amplify existing risks, not only due to their new capabilities, but also due to their unprecedented scale.

The Code's approach to fundamental rights risks makes it an outlier at global level

The AI Act acknowledges international approaches which have identified harmful bias, discrimination, and harms to privacy as fundamental rights risks (Recital 110). Despite this, the third draft applies an interpretation which limits assessment of risks which strictly emerge from high-impact capabilities, leading to the exclusion of fundamental rights risks that the AI Act had already identified as forming part of the emerging consensus on GPAI model risks. For example, the 2025 International AI Safety Report explicitly recognises as systemic risks privacy and environmental risks. In a departure from this increasing consensus, the former risk is relegated to the optional risk category, while the latter is altogether omitted from the draft Code. Consistency with international norms around AI is specifically mentioned in connection with the Code of Practice, with the Act noting that any code "shall take into account international approaches" (Article 56). The current draft's approach to fundamental rights represents a divergence from the consensus-based governance endorsed by the Act.

Lastly, contrary to the conclusion reached by the drafters, fundamental rights risks *do* emerge from high-impact capabilities. Emerging research on dialect prejudice by AI models has found that increasing model scale can make AI models more linguistically prejudiced against African-American English dialect (a dialect associated with descendants of enslaved African Americans in the United States), even as that same increased scale improves a better understanding of that very dialect.

The existing digital rulebook does not cater for the fundamental rights risks emerging from GPAI models in the absence of specific protections in the Code of Practice

The third draft acknowledges the seriousness of the risks including in the optional category, but justifies the changes noting that the they are "better addressed through other parts of the AI Act or other laws, including the Chapters on AI systems in the AI Act, as well as the General Data Protection Regulation (GDPR), Digital Services Act (DSA), and Digital Markets Act (DMA)". While these digital rules have an important role to play in the governance of GPAI models, they do not fully cater to the full range of risks posed by these models.

Let's start with the AI Act. The bulk of the AI Act obligations – and particularly those that focus on fundamental rights – apply to AI systems as opposed to AI models, a distinction that is well understood by the AI Office. The only situation in which these obligations are likely to apply to models is when a GPAI model is integrated into an AI system, which is a distinct possibility envisaged by the AI Act. Such integration, however, does not guarantee that the obligations imposed on AI system providers will apply.

First, the integration of a GPAI model with systemic risks into an AI system does not necessarily mean that the AI system relying on the model will come within the scope of the AI Act. The assessment of whether an AI system poses a risk under the Act – and if so, which type – is entirely separate from the assessment of the risk posed by a GPAI model. It is perfectly possible for an AI system to integrate a GPAI model with systemic risk and to be outside of the scope of the AI Act because the system itself does not pose a transparency risk or a high risk, in which case the AI Act would have nothing to offer by way of additional protections.

Second, even if generic fundamental rights protections are effectively applicable by virtue of the AI system falling into one of the high-risk categories identified in the Act, or otherwise being identified as posing a narrow transparency risk, the Act does not automatically require any type of fundamental rights risk assessment or mitigation, except for the risk management obligations and the – much narrower – obligation to conduct a fundamental rights impact assessment. Both of these avenues have limits. Fundamental rights impact assessments only apply to public authorities and a subset of private entities based on the type of service provided, excluding the vast majority of entities, and the risk management obligation – while it expressly requires consideration of fundamental rights risks – only applies to those risks "which may be reasonably mitigated or eliminated through the development or design of the high-risk AI system or the provision of adequate technical information". If the risk stems from the GPAI model, instead of the high-risk AI system itself, it will not be captured by the provision as ultimately the development or the re-design of the system is not likely to affect the underlying model (Article 9).

Other parts of the AI Act don't set out to protect fundamental rights at large, but instead focus on specific rights. For example, providers of high-risk AI systems are subject to data governance obligations which include the requirement to undertake an assessment of biases likely to have an impact on fundamental rights or lead to discrimination (Article 10). This provision only bites, however, if the provider of the underlying GPAI model is also – by operation of the relevant AI Act sections – the provider of the high-risk AI system.

The practical consequence of the Code of Practice assuming that the AI Act deals with these issues is to offload fundamental rights considerations to AI systems providers and deployers, whenever that system relies on a GPAI model. This interpretation runs counter to the AI Act, which explicitly acknowledges the particular role and responsibility of GPAI models along the value chain, as models that may form the basis for several downstream systems (Recital 101). But it's also unhelpful, considering the limits on documentation-sharing obligations imposed by the Code of Practice itself, which requires limited information to be shared by default with downstream providers. For example, the current draft – consistent with previous versions – requires minimal information on the data used for training and testing to be shared with downstream providers. While it is a positive that the draft code requires the GPAI model providers to document measures to detect unsuitability of data sources, including personal data or harmful content such as CSAM or NCII, as well as measures to detect identifiable biases, disappointingly none of those measures must be disclosed with downstream providers. As a result, the effect of the Code as currently written is to place the entirety of the burden of identifying and mitigating fundamental rights risks on downstream providers. The draft Code includes a catch-call measure whereby GPAI model providers commit to share additional information "necessary to enable downstream providers to have a good understanding of the capabilities and limitations of the GPAI model" (current Measure I.1.2), but it is doubtful that they will actually do so if not actually compelled by the Code.

The draft's reliance on the General Data Protection Regulation (GDPR) – presumably to cover privacy risks – similarly deflects the issue that the Code of Practice was intended to address. While several obligations created under GDPR – ranging from the obligation to comply with the data minimisation principle to the obligation to conduct a data protection impact assessment – are relevant to AI models, their practical application is still the subject of discussion. Last year's European Data Protection Board's opinion on the GDPR and AI models set out the criteria for an AI model to fall within the scope of GDPR, as well as setting out possible mitigations which could have the effect of anonymising a model, and therefore removing it from the law's scope. However, that opinion also made clear that the assessment of any given model's anonymity would be carried out on a case-by-case basis, and acknowledged that the mitigations offered in the opinion were non-prescriptive and non-exhaustive, leaving it to the providers to make their own assessment as to which measures would be necessary to comply with GDPR and which would simply be desirable. The Code of Practice can play a crucial role in clarifying and

standardising approaches by GPAI model providers to the mitigations laid out by the EDPB, boosting enforcement of GDPR instead of treating it as a separate issue.

The draft code takes some steps to protect data protection, notably by requiring GPAI model providers to generate documentation on measures taken to address the prevalence of personal data among the training data, "where relevant and applicable". But nothing in the draft currently requires model providers to proactively take steps to minimise the use of personal data, or to hold them accountable if they do not. Inclusion in the systemic risk taxonomy is a unique opportunity to compel providers to do so.

Two other legal frameworks are referred to as addressing some of the risks excluded from the "selected" systemic risks list: the Digital Services Act (DSA) and the Digital Markets Act (DMA). But these legal instruments only have limited applicability to AI models. While the DSA imposes risk assessment obligations which expressly include risks to fundamental rights, these obligations only apply to very large online platforms and very large online search engines– that is platforms and services that have 45 million EU average active monthly users. In order to be captured by these obligations, a GPAI model provider would not only need to meet this threshold, but more importantly would need to fulfil a dual role, both as a developer of a GPAI model and as a host of an online platform or search engine, which would rely on those models in the design, functioning or use of these services. Not all GPAI models with systemic risks will meet these requirements, limiting the ability of the DSA to be applied as a cross-cutting instrument ensuring fundamental rights safeguards. Further, there is the possibility that a DSA-compliant assessment does not go far enough for the purposes of the AI Act. This is recognised by the AI Act itself, which envisages the possibility for there to be systemic risks not covered by the DSA to emerge (Recital 118). An initial analysis of the first round of Risk Assessment Reports published under Art. 42 of the DSA warns that platforms have focused primarily on user-generated risks, at the expense of risks stemming from the design of their services, including their algorithmic systems. This further highlights the importance of the Code of Practice in independently addressing fundamental rights risks posed by GPAI models with systemic risk.

The DMA poses even further challenges. In order to be captured by the DMA, providers of GPAI models would need to be designated as gatekeepers under the DMA framework. However, while the current list of gatekeepers includes several providers of GPAI models, not all are covered. Once designated as a gatekeeper, the DMA regulates an entity's obligations with regard to their core platform services. While it is notable that the High Level Group for the Digital Markets Act stated that gatekeepers must comply with the DMA's obligations when deploying or embedding AI in their core platform services, this approach shows that the onus of compliance under the DMA at the moment is not on standalone GPAI model providers but on deployers of AI applications into existing core platform services. This is also reflected in emerging literature, which highlights that no consensus about the applicability of the DMA in its current form to GPAI

models exists. Even if GPAI model providers were to fall within the scope of the DMA, the text does not include any form of fundamental rights risk assessment or monitoring obligations for gatekeepers. Moreover, the DMA does not contain any fundamental rights-specific obligations or language. Rather, the High Level Group for the Digital Markets Act acknowledged that risks related to fundamental rights would be relevant for a cross-regulatory discussion.

The positive developments in the Code of Practice do little to address fundamental rights risks

A core positive aspect attaching to risks included in the "selected" risks category is that GPAI model providers must define risk tiers for each of those selected risks, and identify an unacceptable risk tier (Measure II.1.2) for each of those risks. With this requirement, the Code effectively requires model providers to set out red lines which should not be crossed for each of the risks identified by the Code as being mandatory for assessment. The third draft improves on its previous version by now requiring model providers to identify existing processes to facilitate the decision not to release or use a model where contemplated mitigations are insufficient, effectively forcing model providers to consider pressing the red button and withhold the most harmful systems from entering the market. This safeguard would be crucial for fundamental rights risks - but it will not apply to these risks so long as the risks are confined to the optional category in Appendix 1.2.

Another key positive of the current version of the Code is that it strengthens a requirement for signatories who are providers of GPAI models with systemic risk to obtain independent external systemic risk assessments along the model lifecycle (Commitment II.11), starting with a first external assessment prior to making a model available in the market. This commitment introduces a welcome additional layer of scrutiny and accountability which should be preserved moving forward. However, even this positive development does little to address the absence of compulsory fundamental rights considerations in the Code as it currently stands: external assessors are only required to assess the systemic risks *identified* by GPAI model providers. Nothing in the Code compels external assessors to consider omissions or errors in the selection of risks conducted by GPAI model providers (Measure II.11.1), which is an unfortunate limitation on the scope of their work.

If not addressed in the Code of Practice, assessment of these risks will likely fall through the cracks

The issue is not simply that assessment of fundamental rights is optional under the current draft of the Code – but that the Code actively dissuades providers from assessing them by instructing providers to consider these risks where they are reasonably foreseeable, and to "select" them for further assessment only if they are "specific to the high impact capabilities" of GPAI models with

systemic risk. This limitation on the obligation of GPAI model providers to consider risks other than those flagged as compulsory is not only unnecessary, but harmful: it encourages providers to sidestep consideration of fundamental rights risks, just as the Code currently does.

Through these changes, the Code has removed all incentives for providers to account for risks to fundamental rights, leaving it to industry to decide to what extent they assess those risks, if at all.

Conclusion: the third draft presents a step backwards

The downgrading of the fundamental rights risks in the Code of Practice must be seen in light of other changes made to the third draft, including newly introduced changes to public transparency.

Whereas the previous draft encouraged GPAI model providers to publish relevant documents produced under the Code of Practice – namely model frameworks and model reports – allowing for redactions in order to prevent the increase of systemic risks or otherwise avoid divulging commercially sensitive information, the current draft has gone entirely in the opposite direction, requiring publication where *necessary* to effectively enable assessment and mitigation of systemic risks. This makes non-disclosure the default, and information-sharing the exception.

The current version of the Code leaves fundamental rights concerns in regulatory limbo, placing the burden of GPAI model regulation in this regard on existing laws governing the digital space, while paradoxically avoiding addressing the issues specific to the technologies which are at the very core of the Code. Coupled with the limited public insights into GPAI model providers' approach to systemic risks, this creates a dangerous environment where assessment of fundamental rights risks in GPAI models are altogether ignored on the assumption that other entities or applicable frameworks address them. The rights-based framework that the AI Act aimed to create around AI models is severely undermined by the evasive approach undertaken by the draft Code, and will altogether be sacrificed unless major changes are made.