ALLAI.

#Trustworthy GPAI

21.03.2025

AI Act future proof

#3 Third Draft General-Purpose AI Code of Practice

Policy Notes & Feedback

This document serves as input to the Third Draft General-Purpose AI Code of Practice and provides insights to policy makers. It is part of a series of policy documents to inform the AI Act's upcoming guidelines, codes of conduct, implementing and delegated acts.

Authors:
Catelijne Muller, LL.M
Alice Teilhard De Chardin

1. Introduction

This document offers insights on the third and final draft of the General Purpose AI Code of Practice (COP). The input focuses on identifying remaining or new areas of improvement, and highlights critical omissions particularly from the perspective of insufficient alignment with the AI Act and the protection of fundamental rights. It should be noted that this paper does not intend to be exhaustive but merely serves to provide initial input that can be supplemented in the future. We encourage readers to also read our insights on previous drafts of and consultations for the COP. [1]

2. Fundamental Rights & The Systemic Risk Taxonomy

The COP fails to sufficiently safeguard fundamental rights by creating a distinction in Appendix 1 as regards what constitutes a systemic risk:

1. "Selected Types of Systemic Risk"; and
2. "Other Types of Risks for **Potential** Consideration"

This distinction is problematic because the COP only requires identification and mitigations of the systemic risks outlined in category 1: "Selected Types of Systemic Risks". Yet, risks to fundamental rights, society as a whole, public health, safety, and public security are relegated to the second category: "Other Types of Risks for **Potential** Consideration"

The COP justifies this separation by stating it focuses on risks "*specific to the high-impact capabilities of general-purpose AI Models".* Here the COP relies on a flawed interpretation of Article 3(65) of the AI Act. The article defines systemic risk as:

> *"A risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach,* ***or*** *due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or society as a whole, that can be propagated at scale across the value chain."*

The COP's justification for the above separation suggests that only risks stemming from a GPAI model's high-impact capabilities qualify as systemic risks. However, Article 3(65) does not impose such a limitation but rather defines systemic risk in broader terms including

[1] Muller C., Teilhard de Chardin A. (2024): AIA future proof #1 | Trustworthy GPAI Codes of Practice; Muller C., Teilhard de Chardin A. (2024): AIA future proof # | First Draft General Purpose AI Code of Practice (feedback)

significant impact and negative effects including widespread economic and societal effects, negative impacts on fundamental rights, and risks that can spread at scale through applications, integrations, and adaptations.

As such, even if a violation of fundamental rights is not caused directly by the model's high-impact capabilities, it may still spread widely through the model's applications, integrations, or adaptations, making it a systemic risk that should be addressed at the GPAI model level. Recital 110 of the AI Act reinforces this interpretation by stating all systemic risks should be treated with equal weight.

The reasoning behind the COP's interpretation of Article 3(65) is also flawed because systemic risks (including those listed in the Selected Types of Systemic Risk section) are rarely caused solely by the high-impact capabilities of a GPAI model itself. GPAI models exist within a broader socio-technical system, interconnected with people, processes, and networks. By treating systemic risks as if they arise solely from technical capabilities, the COP asserts a false dichotomy that separates AI risks from their real-world context. This narrow framing does not account for the fact that systemic risks are shaped by how AI models are deployed, governed, and integrated into society, rather than merely their inherent capabilities.

We were additionally disappointed by the regression in the COP's systemic risk taxonomy. In the previous draft "large-scale, illegal discrimination" was included in the first category: *Selected Systemic Risk*. In this draft however, "large-scale, illegal discrimination" has been all together removed from the taxonomy.

We see the above elements as representing a weakening of the COP's commitment to addressing fundamental rights violations as systemic risks.

The COP's narrow recognition of systemic risks thus undermines the effectiveness of its commitments on systemic risk identification (Commitment II.3) and systemic risk mitigation (Commitments II.6 and II.7). By failing to account for those broader societal and contextual factors, the commitments ultimately fail to align with the AI Act's emphasis on fundamental rights and broader societal well-being. As a result, the COP's approach risks being too narrow to adequately address the real-world impacts of AI.

The COP only requires identification and mitigations of the systemic risks outlined in category 1: "Selected Types of Systemic Risks", hence a wide range of known and well-documented risks, including those that directly impact individuals daily, such as bias and discrimination, would not require identification or mitigation from signatories.

Finally, the COP''s systemic risk taxonomy does not reflect the full spectrum of systemic risks identified in existing research on the subject. This gap further limits the COP's ability to effectively address real-world harms. Notably, the taxonomy omits risks recognized in the literature, including but not limited to [2],[3]:

- The displacement of humans in the workplace and deterioration of skills
- Model concentration and homogeneity
- Large scale unfair competition
- GPAI providers becoming 'too big to fail'
- Loss of human agency and autonomy
- Broad impact on (a) fundamental right(s)
- The erosion of public trust in social/political institutions
- The perpetuation or exacerbation of inequalities and biases (not merely large-scale illegal discrimination)
- Economic disruptions ranging from large impacts on the labour market to broader economic changes that could lead to exacerbated wealth inequality, instability in the financial system, labour exploitation or other economic dimensions.
- Profound negative long-term changes to social structures, cultural norms, and human relationships that may be difficult or impossible to reverse
- The concentration of military, economic, or political power of entities in possession or control of AI or AI-enabled technologies
- The flooding of information systems with false, spam-filled, or manipulative content, making it hard to discern truth, humanity, and trustworthiness
- Runaway hyper-capitalism arising from largely AI-run companies competing electronically beyond human control or governance.

In sum, the taxonomy and its associated commitments fail to uphold EU values, protect fundamental rights, and ultimately undermine the COP's effectiveness in ensuring the trustworthiness of GPAIs on the EU market.

[2] Ibid.
[3] Uuk, R., Gutierrez, C.I., Guppy, D., Lauwaert, L., Kasirzadeh, A., Velasco, L., Slattery, P. and Prunkl, C., 2024. A Taxonomy of Systemic Risks from General-Purpose AI. arXiv preprint arXiv:2412.07780.

3. ‘Severity’ of Serious Incidents & Systemic Risks

In multiple instances (including *Recital (g)*, *Commitment II.4 on Systemic Risk Analysis*, *Commitment II.12 on Serious Incident Reporting*, and *Measure II.12.2 on Serious Incident Tracking, Documentation, and Reporting*) the COP outlines how systemic risks and serious incidents should be mitigated and reported based on their "severity" or in a manner that is "relative" or "proportionate" to their severity. We take issue with the inclusion of "severity" as a criterion for determining the extent of documentation and mitigation measures required from signatories. All systemic risks and serious incidents should be treated with the full gravity they warrant and addressed with uniform rigor. Doing otherwise allows for dilution based on a subjective assessment of proportionality/relativity.

This approach is also problematic because the introduction of a "severity" criterion risks establishing a differentiated standard of risk management. This differentiation could create an implicit additional risk tier that categorizes systemic risks and serious incidents as "severe" or "non-severe." We oppose this potentiality because detailed risk assessments, mitigations, and documentation should be required whenever a GPAI model poses systemic risks or causes a serious incident regardless of subjective determinations of severity. Indeed, the purpose of the COP is to guide GPAI model providers with a framework for compliance with the AI Act, not to introduce new risk categorizations that fall outside the scope of the AI Act, such as distinguishing between "severe" and "non-severe" systemic risks or serious incidents. This misinterpretation weakens the intent of the AI Act vis à vis the COP. In addition, proportionality, as articulated in the AI Act’s recitals, relates to the burden of compliance on smaller organizations and not to the creation of additional risk levels. As outlined in Recital 10, it should remain commensurate with the capacity of an organization, without compromising the standards of systemic risk management.

4. ‘Safely-Derived Models’

The third draft of the COP introduces a new category of GPAI models, referred to as “safely-derived models” or models deemed “similarly safe”, which are subject to less stringent safety requirements. According to the COP, a model is considered "similarly safe" if it is “as safe as or safer than another GPAISR that has been made available on the market, has undergone systemic risk management as per this Code and meets the systemic risk acceptance criteria of the relevant signatory, or has been made available on the Union market before 2 May 2025.” A "safely-derived model" is one that is “derived directly from the safe originator model” through techniques such as distillation, quantization, fine-tuning, or post-training, or through modifications that solely improve safety or security mitigations.

GPAI model providers with these types of models are exempt from model-specific adequacy assessments (Measure II.9.2), independent external assessments (Commitment II.11), and certain pre-market placement assessments (Measure II.11.1).

We oppose the subclassification and the associated exemptions for various reasons. First and foremost, introducing yet another category of GPAI models goes beyond the boundaries of the AI Act, creating potential regulatory loopholes.

Secondly, the new category of GPAI models and its associated exemptions rely on the flawed assumption that GPAI safety can be conclusively and verifiably established. This assumption incorrectly presumes an originator model can be definitively deemed safe and that the risk assessments and mitigations applied to it were both sufficient and comprehensive. Given the still nascent field of AI safety evaluations and techniques on offer, this can simply not be guaranteed.

Thirdly, by the exemptions rest on the flawed premise that the safety of an originator model will carry over to its derived versions, despite the fact that even minor modifications can introduce new safety risks.[4] Research has shown that safety properties are not always preserved during processes such as fine-tuning, compression, and other forms of model adaptation. For example, quantization can unexpectedly alter a model's behaviour, making LLMs more susceptible to adversarial prompts and unsafe outputs.[5] Similarly, distilled models (which aim to retain the original model's capabilities in a compressed form) have been found to exhibit weaker safety performance compared to their originator models. [6]

Given these, the assumption that a GPAI model will remain "safe" because it is derived from a "safe" originator model is not reliable, especially when it comes to the potential systemic risks at stake. As such, GPAI model providers deploying derived or 'similarly safe' models should not be held to a less rigorous safety standard. Even if a GPAI model appears to pose a low likelihood of systemic risk or is comparable to other established models, such assumptions do not guarantee safety.

---

[4] Qi, X. *et al.* (2023) 'Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!', *arXiv* [Preprint], doi:10.48550/arXiv.2310.03693. http://arxiv.org/abs/2310.03693

[5] Kumar, D. *et al.* (2024) 'Increased LLM Vulnerabilities from Fine-tuning and Quantization', *arXiv.org* [Preprint], doi:10.48550/arXiv.2404.04392. http://arxiv.org/abs/2404.04392

[6] *Distilled, but dangerous? assessing the safety of models derived from DeepSeek-R1* (no date) *Repello AI*. Available at: https://repello.ai/blog/distilled-but-dangerous-assessing-the-safety-of-models-derived-from-deepseek-r1 (Accessed: 28 March 2025).

The absence of systemic harm from a similar model does not mean a newly deployed model will not cause systemic risks, particularly as different contextual factors emerge.

Perhaps most importantly though, the focus on 'safety' ignores the systemic risk GPAI models could pose to fundamental rights and society at large. GPAI models exist within a broader socio-technical system, interconnected with people, processes, and networks, which makes them, even if supposedly "safely-derived", still susceptible to systemic risks. By again treating systemic risks as if they only flow from safety related risks, the COP creates a false dichotomy, failing to recognize that systemic risks emerge from the interaction between AI models and their broader context.

5. Misalignment with the Precautionary Principle

Recital (H) requires GPAI model providers to:

> "*recognise the important role of the Precautionary Principle (laid down in Article 191 TFEU), especially for systemic risks where the lack or quality of scientific data does not yet permit a complete assessment, and will take the extrapolation of current adoption rates and research and development trajectories of GPAISRs into account for the identification of systemic risks".*

Article 191 of the TFEU specifically outlines the need for "*preventative decision-taking in the case of risk"* when the following conditions are met: identification of potentially adverse effects; evaluation of the scientific data available; the extent of scientific uncertainty.

The COP's recognition of the Precautionary Principle appears inadequate when assessed against its requirements for GPAI model safety and security. While the COP acknowledges the Precautionary Principle's importance, particularly for risks where scientific data is insufficient for complete assessment, it insufficiently requires preventative action or proactive efforts to actively reduce systemic risks. Instead, it requires risks be mitigated to remain below the "unacceptable risk" tier, which we argue reflects an overly permissive interpretation of the Precautionary Principle.

In its current form, the COP relies on the "unacceptable risk" criteria, requiring GPAI model providers to identify **at least one unacceptable tier of risk** for each selected systemic risk type the GPAI model may pose. It is only when a model is projected to reach this tier and when no effective mitigations exist/fail that the COP explicitly recommends stopping development.

Here the COP introduces a categorization to the concept of systemic risk, introducing levels of systemic risks that are supposedly acceptable. We take issue with this approach as we strongly believe that no systemic risk should be considered acceptable.

Also, this approach introduces an issue in those cases where data or evaluations on a GPAI model's risk level are incomplete. In such cases, the COP advises GPAI model providers to reasonably foresee potential risks by extrapolating trends or consulting independent experts. Yet, the COP still relies on the GPAI model provider's own risk assessment to determine a risk unacceptable. This reliance on subjective assessments rather than a clear requirement to stop development when evidence is inconclusive/in cases of uncertainty, increases the risk of delayed preventative action. If GPAI model providers take an overly optimistic view when faced with uncertainty or if unforeseen issues emerge (which they often do), the result could be avoidable harm that was not adequately prevented.

6. 'Systemic Risk Acceptance'

The COP adopts a risk threshold approach, categorizing systemic risks as either "acceptable" or "unacceptable". Under the *Security and Safety Framework*, GPAI model providers must establish systemic risk acceptance criteria defining what level of systemic risk is deemed acceptable or unacceptable vis a vis their GPAI model. They then compare the results of their risk analysis to these criteria to determine whether they can proceed with developing or deploying their GPAI model on the EU market. If the estimated risk falls below the "unacceptable" threshold and remains within the "acceptable" range, development may continue with standard mitigations. However, if the risk is found to exceed this threshold (deemed unacceptable), the COP requires stronger measures such as additional mitigations or the reconsideration of deployment.

We take issue with this conceptualization of risk thresholds because it again introduces a categorization of systemic risks, allowing for some systemic risks to be deemed "acceptable". By definition, a systemic risk represents a potential for significant impact on both the Union market, public health and safety, fundamental rights and/or society as a whole. Therefore, there can be no such thing as an "acceptable" significant impact, nor an "acceptable" level of systemic risk that threatens market stability or public well-being.

By framing systemic risks as "acceptable", the COP raises legal ambiguities that should be avoided for alignment with the AI Act. Introducing risk acceptance criteria and thresholds, creates subjective room for interpretation regarding what qualifies as "acceptable" systemic risk.

This could lead to inconsistent standards because individual GPAI model providers set their own thresholds, potentially leading to varying degrees of tolerance for harm and systemic risk. In addition to creating legal uncertainty and room for circumvention of the AI Act, this approach is also inconsistent with the Precautionary Principle, which mandates minimizing and reducing systemic risk as much as possible, rather than merely keeping risk within an arbitrary threshold.

Therefore, the COP should reconsider its approach to systemic risk and the way GPAI model providers address the risks posed by their GPAI models. Instead of mitigating systemic risks to an "acceptable level," the COP should promote a proactive risk management approach that prioritizes systemic risk reduction and minimization. This shift would ensure GPAI model providers actively work to eliminate or significantly reduce systemic risks.

In addition, the COP should revise its language and terminology. We believe the current "acceptable" vs. "unacceptable" distinction is unhelpful and could lead to interpretive ambiguity. Furthermore, the language of “acceptable” systemic risks could foster a dangerous narrative that implicitly suggests some level of systemic risk is indeed acceptable/tolerable. We maintain the default position should be that all systemic risks are inherently unacceptable. Therefore, a conceptual and terminological shift would better align the COP with a precautionary approach that is additionally aligned with the protection of EU values.

7. Safety Mitigations and Measures

Measure II.6.1 requires GPAI model providers to:

> “*as necessary, mitigate systemic risks and reduce them to acceptable levels (in accordance with this Code), implement technical safety mitigations for GPAISRs, such as: (1) filtering and cleaning training data; (2) monitoring and filtering the inputs and outputs of such models; (3) changing the behaviour of such a model in the interests of safety, such as finetuning the model to refuse certain requests; (4) restricting the availability of such a model on the market, such as restricting model access to vetted users; (5) offering countermeasures or other safety tools to other actors; (6) implementing high-assurance quantitative safety guarantees concerning the behaviour of such a model; and (7) implementing infrastructure that could help promote safe ecosystems of AI agents, such as a reputation system, specialised communication protocols, or incident monitoring tools.”.*

Each of these seven safety mitigations to implement in Measure II.6.1 play a role in mitigating systemic risks from GPAI models. However, we emphasize that none of these measures alone can sufficiently reduce or mitigate systemic risk. Therefore, the COP should be supplemented with an explicit statement clarifying the effectiveness of an overall mitigation strategy will hinge on the utilization of these mitigations in combination, along with a commitment to continually improving each (as outlined in *Recital (a)*).

Furthermore, we propose additional mitigations to be included in the list, such as the use of transparency and interpretability tools, differential privacy techniques, and the development of ongoing human-in-the-loop oversight.

The latter is particularly important given that the COP suggests the use of "*automated benchmarks enabling highly scalable and real-time identification of capability increases*" as a potential procedure for detecting changes in systemic risks that may necessitate pausing GPAI development for further risk assessment in Measure II.2.2. On this point, we believe an additional sentence should be added to this measure to emphasize that automated benchmarks should never replace human oversight and human applied benchmarks. This is important because automated benchmarks can fail, either due to technical issues or misalignment, but also because relying solely on them can result in a loss of domain understanding by human evaluators.[7] When oversight is outsourced to AI, layers of AI evaluation can make it difficult for the human evaluator to pinpoint where issues arise, leading to opacity in the evaluation process.

Measure II.6.2. requires GPAI model providers to:

> "*implement state-of-the-art technical safety mitigations that: (1) are proportionate to the systemic risk at issue, taking into account the systemic risk acceptance criteria (as set out in the Framework pursuant to Measure II.1.2); and (2) best mitigate, in particular, unacceptable systemic risks (as set out in the Framework pursuant to Measure II.1.2). For the purposes of this Code, state-of-the-art technical safety mitigations need not always or necessarily mitigate all systemic risks to the greatest extent".*

[7] Bowman, S. *et al.* (2022) *Measuring Progress on Scalable Oversight for Large Language Models* [Preprint]. doi:10.48550/arXiv.2211.03540. http://arxiv.org/abs/2211.03540

We argue the concluding sentence of Measure II.6.2 should be removed due to the potential misinterpretation or strategic manipulation of the qualifier "state-of-the-art." First, this language could be interpreted as allowing GPAI model providers to implement only partial, superficial, or minimal mitigations, as long as they meet the "state-of-the-art" standard. This would dilute the AI ACT' intent by implying it is acceptable for GPAI model providers to not strive for full or near-complete risk mitigation. Second, the phrasing may introduce a loophole, enabling GPAI model providers to claim adherence with the COP by adopting a "state-of-the-art" mitigation technique, even if it only addresses part of the systemic risk.

Measure II.6.3 requires GPAI model providers to:

> "*implement corrective measures to address serious incidents through the use of heightened or additional technical safety and/or security mitigations (pursuant to this Commitment and/or Commitment II.7)*".

Measure II.6.3 has been revised from the second draft of the COP to remove the requirement for GPAI model providers to pre-define corrective measures, now only requiring providers implement corrective measures in response to serious incidents. We disagree with this change, as it weakens *incident response readiness* (which is important given the measure's title). While we acknowledge the complexity of real-world incidents may mean pre-defined measures are not always fully applicable, having a set of pre-established corrective actions ensures GPAI model providers are better prepared to respond quickly and effectively in the event of an incident. Without this requirement, GPAI model providers may insufficiently consider potential serious incident scenarios which could lead to delayed or inadequate responses when incidents occur.

8. Whistleblower Protections

Commitment II.13 requires GPAI model providers to:

> "*commit to not retaliating against any worker providing information about systemic risks stemming from the AI model providers' GPAISRs to the AI Office or, as appropriate, to national competent authorities, and to at least annually informing workers of an AI Office mailbox designated for receiving such information, if such a mailbox exists*".

We take issue with Commitment II.13 as it has removed most of the whistleblower protections which were incorporated in the second draft of the COP. The previous draft

featured a dedicated commitment titled, "*Commitment 18. Whistleblowing protections"* which explicitly mandated adherence to Directive (EU)2019/1937 on the protection of whistleblowers and implementing laws of Member States. Resultantly, the previous commitment included measures such as those requiring GPAI model providers to commit "*to implementing whistleblowing channels and afford appropriate whistleblowing protections to covered persons and activities, as per Directive (EU) 2019/1937"*.

Meanwhile, the latest draft of the COP merely requires GPAI model providers commit to non-retaliatory measures in response to whistleblowers, while the reference to the Directive has been relegated to the recital rather than explicitly included in the commitment itself. We argue this change represents a backsliding in the protection of whistleblowers. Instead, whistleblower protections should be bolstered by requiring explicit adherence to the Directive in the COP.

9. Public transparency

Commitment II.16. requires GPAI model providers to:

> "*commit to publishing information relevant to the public understanding of systemic risks stemming from their GPAISRs, where necessary to effectively enable assessment and mitigation of systemic risks, to the extent and under the conditions specified in Measure II.16.1"*.

Measure II.16.1 requires GPAI model providers to publish their Safety and Security Framework, along with the Model Report (or equivalent documentation), to enhance public awareness of systemic risks associated with a model, strengthen societal resilience against these risks, and support the detection of systemic risks. While we commend the COP's requirement for GPAI providers to publish their Framework alongside their Model Report, we argue that to promote genuine and meaningful transparency, providers should also disseminate simplified, layperson-friendly companion documents. These documents should explain the implications of the Model Report and Framework in clear and accessible language for the general public.

Furthermore, the Reports and Frameworks should be made available through appropriate public channels and integrated into the product in a consumer-facing manner. For example, users should be informed about key risks, such as a 20% chance of model leakage, alongside an explanation of the potential negative implications. At the same time, providers should highlight the specific mitigation efforts they have undertaken to reduce such risks and

demonstrate the resulting improvements. This approach would ensure the public is not only aware of the risks but also sees the proactive steps being taken to address them, fostering greater trust in GPAI models on the EU market.

10. The COP Drafting Process

We would like to raise final concerns regarding the COP drafting process. While the process has been promoted as inclusive, involving contributions from over a thousand stakeholders from academia, civil society, EU member state representatives, international observers and industry over several months, it is disappointing to see that successive drafts of the COP have not meaningfully incorporated the voices of these diverse stakeholders, particularly those from civil society and fundamental rights experts.

Having actively participated in multiple COP Working Groups and provided extensive survey feedback on the drafts (continuously emphasizing and substantiating the need for stronger protections of fundamental rights) we are concerned this input, along with similar feedback from many others, has not been subsequently integrated into the COP.

With this being the final round of open stakeholder feedback before a closed-door workshop with GPAI model providers from industry (scheduled for March 28th), we question whether this workshop will water down the limited protections of fundamental rights, democracy and society at large even further.

All this raises concerns that the wider stakeholder engagement process could ultimately amount to little more than window dressing undermining the claim of the COP being a representative and balanced piece. As a result, we urge the chairs of the COP to reflect on and acknowledge the insufficient integration of civil society and fundamental rights perspectives. It is important to take meaningful steps to address this issue in these final stages of the drafting process to uphold the integrity of the EU market and safeguard the core principles of accountability, fundamental rights protection, and democracy that lie at the heart of the European Union.

ABOUT ALLAI

ALLAI is an independent organisation that aims to foster, promote and achieve the responsible development, deployment and use of AI.

ALLAI's mission is to take a holistic approach to AI, taking into account all impact domains such as economics, ethics, privacy, laws, safety, labour, education, etc. ALLAI aims to involve all stakeholders in its mission: policy-makers, industry, social partners, consumers, NGOs, educational and care instructions, academics from various disciplines.

ALLAI was founded by the three Dutch members of the High Level Expert Group on AI, Catelijne Muller, LLM, Prof. Virginia Dignum and Associate Prof. Aimee van Wynsberghe. Collectively, the founders have a broad expertise in AI: AI sciences, social impact, national and international policy, legal implications, and ethical impact.

CONTACT

ALLAI
Amsterdam Science Park
LAB42
The Netherlands

www.allai.nl
@ALLAI_EU
welkom@allai.nl