ALLAI.

#Trustworthy GPAI

27.11.2024

AIA future proof

#2 First Draft General-Purpose AI Code of Practice

Feedback

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

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

1. Introduction

This document offers supplementary input for the Multi-Stakeholder Consultation FUTURE-PROOF AI ACT: TRUSTWORTHY GENERAL PURPOSE AI. Specifically, it addresses key considerations relevant to Working Groups WG2, WG3, and WG4. The input focuses on identifying **areas of improvement in the initial draft of the GPAI Code of Practice (COP)** and highlights critical omissions that merit further evaluation for potential inclusion in the final COP.

It should be kindly noted that this paper does not intend to be exhaustive, but merely serves to provide initial input that can be supplemented in the future.

2. Title and High Level Principles - Trustworthiness

In the first consultation for the COP, carried the title: *A Codes of Practice for **Trustworthy** GPAI*. The first draft of the GPAI COP does no longer mention the word 'Trustworthy'. Moreover, in the Chapter 'Drafting plan and principles under point **I. Alignment with EU Principles and Values**, reference is made to EU Law only, and not to any (other) principles and values. In 2019, the EU High Level Expert Group on AI set 'Trustworthiness' as an EU benchmark for AI that is lawful, ethically aligned and robust. The AI Act and other regulations deal with the pillar of lawfulness. The two other pillars, **Ethical Alignment** and **Robustness** consist of 7 requirements reflecting key EU principles and values that GPAI should be grounded in as well.

Trustworthiness also involves asking **Question 0** – should this system be developed and deployed in the first place? The COP should require GPAI providers to ask and answer this question. Particularly when a systemic risk is posed by their system, GPAI providers should be obligated to decide not to further develop or even deploy the system.

Recommendation:

- Amend the title to **A Codes of Practice for Trustworthy GPAI**
- Reference the **Ethics Guidelines for Trustworthy AI** and ensure adherence to the 3 pillars of Trustworthy AI as elaborated in the Ethics Guidelines for Trustworthy AI: Lawfulness (incl. adherence to the AI Act, but also other regulations), Ethical alignment and Robustness (as described in the 7 key requirements).

3. GPAI Innovation & Downstream Providers

WG 4
Ref. to II RULES FOR PROVIDERS OF GPAI MODELS, Transparency, Measure 2. Documentation for Down-Stream Providers

Many GPAI models will likely be integrated into 'intended purpose' AI systems categorized as high risk or limited risk AI under Art. 6 of the AI Act. The responsibility of bringing these systems in compliance with the AI Act lies with the 'downstream' providers and deployers of these high risk or limited risk AI systems. They will be the ones having to comply with, for example, the requirements for high risk AI around data quality and data governance (art. 13), transparency (art. 14), human oversight (art. 14) and accuracy (art. 15). As many of these requirements are technical in nature or require certain technical specificities, compliance will often have to be met at GPAI model level.

If there continues to be a lack of alignment between the COP for GPAI models and the requirements for high risk AI, this could lead to a limited uptake of GPAI systems on the one hand, and a (further) concentration of AI innovation power with GPAI developers on the other. Hence, the AI Act's risk based approach for 'intended purpose AI' mandates alignment of the GPAI Codes of Practice with the prohibitions, as well as the requirements for high risk and limited risk AI. Without such alignment there is a serious risk of stifling innovation rather than supporting it.

4. Proportionality

WG 4
Ref. to II RULES FOR PROVIDERS OF GPAI MODELS, point C

The draft COP writes, *"The Signatories recognise that in the case of a modification or fine-tuning of a model, the obligations for providers should be limited to that modification or fine-tuning to safeguard proportionality.'*

The wording of this sentence, that is is related to proportionality, is unclear and could introduce regulatory uncertainty or ambiguity. The current phrasing fails to specify which providers' obligations are limited when their obligations are limited to modification or fine-tuning. Are these the original GPAI providers, downstream GPAI providers or downstream providers that build a system with and intended purpose using a GPAI model? This ambiguity is problematic because downstream providers developing intended-purpose systems would fall under a different regulatory regime with distinct high-risk AI requirements. As such, this

sentence should explicitly reference which specific category of providers it refers to, as well as provide clarity on the precise scope of the requirements related to the modification, and how these limitations interact with the existing AI Act provisions.

5. Taxonomy of Systemic Risks

WG 2
Ref. to III. TAXONOMY OF SYSTEMIC RISKS, section 6.1

Whilst the COP includes an important list of elements which signatories will treat as systemic risks, we argue this list is insufficient.

Recommendation:
Add the following systemic risks:

- Displacement of humans in the workplace and deterioration of skills
- Model concentration and homogeneity
- Environmental impact
- Large scale unfair competition
- GPAI providers becoming 'too big to fail'
- Loss of human agency and autonomy
- Broad impact on (a) fundamental right(s)

Ref. to III. TAXONOMY OF SYSTEMIC RISKS Point C

The draft COP outlines how: *"The Signatories recognise that the taxonomy of systemic risks is non-exhaustive and will be subject to change over time, reflecting scientific advances and societal changes."*

We raise concerns, given the significant societal and technological shifts that have occurred in just the last three months, on how the AI Office intends to keep the taxonomy of systemic risks current and up-to-date amid such rapid and unpredictable changes. The taxonomy should be revised and updated frequently to incorporate relevant developments. We recommend the explicit inclusion of a risk-forecasting methodology which specifically sets out measures to analyse, examine, and update systemic risks in line with contemporary geopolitical, economic, societal, and environmental uncertainty.

Recommendation:
Develop a strategy and methodology to revisit and update the systemic risk taxonomy to future proof the COP through risk-forecasting.

6. Sources of Systemic Risk

WG2

Ref. to III. TAXONOMY OF SYSTEMIC RISKS, Section 6.3.3

We believe the list of systemic risk sources, particularly those related to *model affordances and the socio-technical context*, should include risks associated with model concentration and homogeneity. These factors, which extend beyond the inherent properties of individual models, could significantly amplify the systemic risks posed by a particular model. Today, there is a trend towards ever fewer, ever more 'general', and ever larger models. While these models have demonstrated impressive behaviour, they can also fail unexpectedly (hallucinate), harbour biases, and are poorly understood. As these systems are deployed at scale, they can become singular points of failure that radiate harm (e.g., security risks, discrimination, inequities) to countless downstream AI applications. The multiple legal and ethical issues these models present, such as around data protection, IP rights, automation bias, manipulative power, the scaling of misinformation, skills erosion, potential job displacement, the risk of uncontrollable autonomy, and so on, are well known. If ever fewer machines inform ever more decisions, biases and errors could become amplified and embedded to the point where they create structural societal drawbacks (Creel and Hellman (2021). Hence, for a GPAI model, one could argue that because of their potential use in a wide variety of high-risk domains (healthcare, critical infrastructure, law enforcement), they should be held to a *higher* standard by the AI Act.

Recommendation:
Include model concentration and model homogeneity as sources of systemic risks, promting GPAI models to be held to a higher standard.

7. Systemic Risk Assessments and Mitigation

WG 2

Ref. to IV. RULES FOR PROVIDERS OF GENERAL-PURPOSE AI MODELS WITH SYSTEMIC RISK, Measure 8. Risk identification

We commend the COP to include measures requiring GPAI model providers to continuously identify systemic risks that may stem from their GPAI models. However, we additionally argue that in addressing the domain of systemic risk assessments, the COP must develop a dynamic and responsive framework that ensures evidence-based evaluations can evolve

with emerging technological and societal insight. Indeed, looking at the current risk assessments frameworks, we worry the COP insufficiently ensures that systematic risk assessments remain adaptable as new information comes to light. In doing so, the COP should include provisions for continuous knowledge integration between GPAI model providers, AI labs, expert analysis and broader scientific research to be shared and systematically reviewed and incorporated into risk assessment protocols.

In conjunction with the previous point, the COP should include mechanisms or processes for societal stakeholders and individuals to report sources of systemic risk or incidents. For instance, if a stakeholder experiences changes in model behavior during interactions, such as instances of deception or unexpected persuasive capabilities, there should be a formal process for them to raise these concerns. As such, a framework for systemic risk assessments could also include accessible channels for broader societal stakeholders to report potential systemic risks or significant incidents.

Recommendations:

- The COP should ensure the presence of a dynamic and responsive framework that ensures evidence-based evaluations can evolve with emerging technological and societal insight.
- Accessible channels for broader societal stakeholders to report potential systemic risks or significant incidents should be made available.

WG3

Ref. to "Substantial" Systemic Risks in IV. RULES FOR PROVIDERS OF GENERAL-PURPOSE AI MODELS WITH SYSTEMIC RISK, point B

Point B of the draft COP writes: *"Signatories recognise that detailed risk assessment, mitigations, and documentation are particularly important where the general-purpose AI model with systemic risk is more likely to (i) present substantial systemic risk, (ii) has uncertain capabilities and impacts, or (iii) where the provider lacks relevant expertise"*

The distinction between "systemic risk" and "*substantial* systemic risk" in the draft COP introduces an unnecessary and potentially dangerous semantic nuance. This word effectively creates a new categorization that undermines the nature of systemic risks by implying that some systemic risks might be less consequential. We strongly oppose this distinction because by definition systemic risks are inherently substantial. They represent potential threats that could compromise entire infrastructures, lives, and fundamental rights. These have implications that cannot be stratified into degrees of significance.

We note that reference to “substantial systemic risk” with associated differing obligations are also mentioned in:

GOVERNANCE RISK MITIGATION FOR PROVIDERS OF GENERAL-PURPOSE AI MODELS WITH SYSTEMIC RISK, Measure 20. Notifications: *“Signatories commit to notify the AI Office of relevant information regarding their models meeting the thresholds for general-purpose AI models to classify as general-purpose AI models with systemic risk, their SSF, their SSR, and substantial systemic risks where appropriate”*

Sub-Measure 20.4. Substantial systemic risk notification: “*Signatories will notify the AI Office if they have strong reason to believe substantial systemic risk might Materialise.”*

IV point B: *“Signatories recognise that detailed risk assessment, mitigations, and documentation are particularly important where the general-purpose AI model with systemic risk is more likely to (i) present substantial systemic risk.”*

Recommendation:

Remove any reference to **‘substantial’** systemic risk with associated differing obligations.

WG3

Provision for SMEs (Ref. to IV. RULES FOR PROVIDERS OF GENERAL-PURPOSE AI MODELS WITH SYSTEMIC RISK, point B

The COP outlines, *“Signatories recognise that detailed risk assessment, mitigations, and documentation are particularly important where the general-purpose AI model with systemic risk is more likely to (i) present substantial systemic risk, (ii) has uncertain capabilities and impacts, or (iii) where the provider lacks relevant expertise. To account for differences in available resources between providers of different size and capacity, and recognising the principle of proportionality, simplified ways of compliance for SMEs and startups will be provided where appropriate.”*

Whilst recognising the potential difficulty and challenges for SMEs and startups to adhere to the rules, we argue that reducing safety requirements based on resource constraints could allow (systemic) risks from GPAI models to proliferate unchecked. Hence safety, testing, and mitigation standards must be equally rigorous regardless of the size of the GPAI model provider.

Recommendation:

Given the potential systemic risk of GPAI models, regardless of the size or resources of the actor developing and deploying them, the COP should provide for equally rigorous safety, testing and mitigation requirements for SME's, micro-enterprises and startups.

WG3

Best-in-class Assessments & Evaluations, Ref. to IV. RULES FOR PROVIDERS OF GENERAL-PURPOSE AI MODELS WITH SYSTEMIC RISK, RISK ASSESSMENT FOR PROVIDERS OF GENERAL-PURPOSE AI MODELS WITH SYSTEMIC RISK.

We found the COP draft consistently refers to the requirement for providers to evaluate their models using best-in-class / state-of-the-art evaluations and AI safety techniques:

"They will make use of a range of methods from forecasting to best-in-class evaluations to investigate capabilities, propensities, and other effects of these models."

"Signatories will ensure best-in-class evaluations are run to adequately assess the capabilities and limitations of their general-purpose AI models with systemic risks... using a range of suitable methodologies"

"Signatories will ensure that evaluations are being run with a best-in-class level of capability elicitation"

The references to 'best-in-class' evaluations are vague, underspecified and unhelpful. Indeed, this terminology creates ambiguity, particularly for GPAI model providers who may lack expertise in AI safety methodologies. Without specific reference to detailed frameworks, methodologies, benchmarks or assessment techniques, GPAI model providers may default to minimal or inappropriate evaluation strategies, believing they have met the standard.

In addition, the COP provides no actionable guidance on what constitutes "best-in-class" evaluations. This fault may be a symptom of the larger issue at hand: that AI trustworthiness, including safety, research is significantly lagging and struggling to keep pace with the rapid development of GPAI technology. As such there are currently limited tests, evaluations, and methodologies available to ensure the trustworthiness, including safety, of GPAI models. We argue that the COP needs to address this gap, particularly since the constant evolution of both the AI trustworthiness landscape and AI technologies more generally means that what may be considered "best" today may be obsolete tomorrow.

We also argue the COP needs to ensure that risk mitigation measures for systemic risks are not merely dealt with and seen through the lens of technical AI safety endeavors. We believe the current approach to AI safety risks reduces systemic risk mitigation to a purely technical exercise, which misunderstands the complex and socio-technical nature of GPAI models. As such, AI trustworthiness research should not be confined to technical safety, but must also integrate EU values and fundamental rights considerations. This integration requires a more holistic perspective that recognises AI safety as a mult-dimensional and contextually-adaptive approach where technical measures remain essential, but must be complemented by societal and cultural measures too.

In line with the objective of the AI Act, which is to protect health, safety and fundamental rights from the ill effects of GPAI, this integration also means developing 'best-in-class' evaluations from a socio-technical and fundamental rights based approach.

WG3

Ref. to ISO standards

During the WG3 meeting, multiple references were made to existing ISO standards. While some stakeholders may view a COP that is aligned with ISO standards as adequate, these standards alone are insufficient to address the complex risks posed to fundamental rights, democracy, the rule of law, and society as a whole, as was also clearly concluded in a recent Science for Policy Brief of the Joint Research Centre[1].

Recommendation:

The COP should ensure that GPAI trustworthiness research keeps pace with GPAI development, and is not confined to technical or safety focussed research only, but integrates research into societal impact, and EU values' and fundamental rights' implications. Existing ISO standards do not suffice to address the latter.

[1] SOLER GARRIDO, J., DE NIGRIS, S., BASSANI, E., SANCHEZ, I., EVAS, T., ANDRÉ, A. and BOULANGÉ, T., Harmonised Standards for the European AI Act, European Commission, Seville, 2024, JRC139430.

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