ALLAI.

#Trustworthy GPAI

17.11.2024

AIA future proof

#1 Trustworthy GPAI Codes of Practice

Article 56

This document serves as input to the Multi-stakeholder Consultation: FUTURE PROOF AI ACT: TRUSTWORTHY GENERAL-PURPOSE AI. 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 provides additional input for the Multi-stakeholder Consultation FUTURE-PROOF AI ACT: TRUSTWORTHY GENERAL PURPOSE AI and should be read in connection with the replies provided to the online questionnaire. The additional input in this paper is specifically related to:

- The concept of 'Trustworthy AI' (Chapter 2)
- GPAI trustworthiness and innovation (Chapter 3)
- GPAI in the AI Act (definitions and interaction with Chapters II & and IV) (Chapter 4)
- Systemic Risks (taxonomy, sources, identification and evaluation) (Chapter 5)

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. Trustworthy AI

In its Communication of April 25th, 2018, the European Commission introduced its vision for AI. Three pillars underpinned that vision: (i) increasing public and private investments in AI to boost its uptake, (ii) preparing for socio-economic changes, and (iii) ensuring an appropriate ethical and legal framework to strengthen European values.

Building on this vision, the Commission assigned a diverse group of 52 experts from academia, business and civil society (High Level Expert Group on AI) to advise the Commission and prepare two 'deliverables': 1. Ethical guidelines for AI and 2. Policy and investment recommendations on AI.

2.1 The Ethics Guidelines for Trustworthy AI

The first deliverable was finalised in April 2019: The Ethics Guidelines for Trustworthy AI (the "Guidelines"). Here the notion of "trustworthiness" was first introduced and it is important to recall this notion, because it still remains the fundament upon which AI in Europe sits.

Trustworthiness is strongly rooted European Union values and fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union. It was chosen as the core theme since it is a prerequisite for social and economic benefits of AI to materialize.

The Guidelines emphasize that the development and use of AI needs to be "human-centric", i.e. in the service of humanity and the common good: "We also want producers of AI systems to get a competitive advantage by embedding Trustworthy AI in their products and services. This entails seeking to maximize the benefits of AI systems while at the same time preventing and minimizing their risks."

The Guidelines outline **three core components** that 'trustworthy AI' needs to adhere to:

1. It should be **lawful**, complying with all applicable laws and regulations
2. It should be **ethical**, ensuring adherence to ethical principles and values
3. It should be **robust**, both from a technical and social perspective

Lawful AI

While the Guidelines do not give any recommendations on laws and regulations for AI, they do emphasize that AI systems do not operate in a lawless world and that any development or use of AI systems must adhere to existing laws. Multiple legally binding rules at European, national and international level already apply or are relevant to AI today, as the Guidelines specifically state. Legal sources include, but are not limited to: EU primary law (the Treaties of the European Union and its Charter of Fundamental Rights), EU secondary law (such as the GDPR, the Product Liability Directive, Safety and Health at Work Directives), UN Human Rights treaties and the Council of Europe conventions (such as the European Convention on Human Rights), EU Member State laws and various domain-specific rules.

Ethical AI

Achieving Trustworthy AI requires not only compliance with the law, which is but one of its three components. Laws are not always up to speed with technological developments, can at times be out of step with ethical norms or may simply not be well suited to addressing certain issues. For AI systems to be trustworthy, they should hence also be ethical, ensuring alignment with ethical norms.

Robust AI

Even if an ethical purpose is ensured, individuals and society must also be confident that AI systems will not cause any unintentional harm. AI should perform in a safe, secure and reliable manner, and safeguards should be foreseen to prevent any unintended adverse impacts. This is needed both from a technical perspective and from a social perspective.

The pillars of Ethical and Robust AI where further defined in 7 key requirements for trustworthy AI:

1. Human Agency and Oversight
2. Technical Robustness and Safety
3. Privacy and Data Governance
4. Transparency
5. Diversity, Non-discrimination and Fairness
6. Environmental and Social Well-being
7. Accountability

The Guidelines then identified the technical and non-technical methods to achieve Trustworthy AI: proper system architecture, trustworthiness by design, explanation methods, testing and validation, quality of service indicators, regulation, codes of conduct, standardization and certification, governance, education & awareness, stakeholder participation and social dialogue, diverse and inclusive design teams.

While major advancements have been made since, particularly in the regulatory domain, with the AI Act, the proposal for the AI Liability Directive and the AI Convention of the Council of Europe (to be implemented by the AI Act), ALLAI emphasizes that Trustworthy AI is to be understood as descirbed in the Guidelines. Achieving Trustworthy AI requires not only compliance with the law, which is but one of its three components: ethical and robust. Laws can at times be out of step with ethical norms or may simply not be well suited to addressing certain issues. Moreover, ethical norms can aid in the interpretation of laws. For AI systems to be trustworthy, they should hence also be ethical, ensuring alignment with ethical norms. Even if an ethical purpose is ensured, individuals and society must also be confident that AI systems remain robust and will not cause any unintentional harm.

Recommendation:
A Codes of Practice for Trustworthy GPAI should include and adhere 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 Trustworthiness and Innovation

There is an obvious 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 harms (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.

Apart from ‘general’ AI-models, there is a wide practice of using so-called ‘benchmark’ datasets that form the backbone of machine learning research and development. Critical inquiry into these datasets have however revealed biases, poor categorization and offensive labelling in these datasets. Koch et al. (2021) have found increasing concentration on fewer and fewer datasets in the field of AI research.

The issues described above around GPAI (consisting of ever fewer and more general models and benchmark datasets) can be referred to as the 'homogeneity problem'. Machine learning by its nature results in more homogeneous decision making compared to human decisions. 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 GPAI one could even 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 in stead of a lower one. In fact, the overall Union objective of safety and liability legal frameworks, is to ensure that all products and services, including those integrating emerging digital technologies, operate safely, reliably and consistently and that damage is remedied efficiently. The EU follows a different approach than other parts of the world, where responsibility is determined afterwards, often leading to large liability claims. It would also break with the overall objective of the AI Act which is to protect health, safety and fundamental rights from adverse effects of AI.

Many GPAI models will likely be integrated into high risk or limited risk AI systems. The responsibility of bringing these systems in compliance with the AI Act lies with the 'downstream' providers en deployers. 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), or accuracy (art. 15). As these may of these requirements are technical in nature, compliance can often only be met at GPAI model level. Disalignment of GPAI models and the requirements for high risk and limited risk AI, 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 approacht for 'intended purpose AI' mandates alignment of the GPA 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 in fact stifling innovation rather than supporting it.

Recommendation:
A Codes of Practice for Trustworthy GPAI should be aligned with the prohibitions, as well as the requirements for high risk and limited risk AI of the AI Act, in order to support rather than stifle innovation.

4. GPAI in the AI Act

The AI Act makes a distinction between ‘general purpose AI *models’* and ‘general purpose AI *systems’*.

Art. 3 (63) defines ‘general-purpose AI model’ as: an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market.

Art. 3 (66) ‘general-purpose AI system’ means an AI system which is based on a general-purpose AI model, and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems.

While ‘AI-system’ is defined in art. 3(1), ‘AI model’ is not. Recital 97 does give some guidance by stating that ‘although AI models are essential components of AI systems, they do not constitute AI systems on their own. AI models require the addition of further components, such as for example a user interface, to become AI systems.’

Recital 100 further indicates that ‘when a general-purpose AI model is integrated into or forms part of an AI system, this system should be considered to be general-purpose AI system when, due to this integration, this system has the capability to serve a variety of purposes. A general-purpose AI system can be used directly, or it may be integrated into other AI systems.’

Seen in light of the two ‘regimes’ of the AI Act: (i) the ‘intended purpose regime’ of Chapters II, III and IV) and (ii) ‘general purpose regime’ of Chapter V, and their diverging requirements, it is important to know when an AI model is considered having a specific purpose (and is covered by the ‘intended purpose regime’) or a general purpose (covered by the ‘general purpose regime’).

Recommendation:
The delimitation of when AI models are considered having a specific purpose and when they have AI a general purpose is something A Codes of Practice should provide clarity on, to avoid legal uncertainty as well as circumvention of the rules.

5. Systemic Risks

The AI Act introduces the concept of systemic risk in relation to GPAI models. This chapter will examine the criteria established by the AI Act for assessing systemic risk and analyze the relationship between high-impact capabilities and systemic risk. Additionally, the chapter will highlight potential challenges and ambiguities within this classification and criteria, while offering alternative proposals. It will subsequently include a systemic risk taxonomy, risk identification and assessment measures, technical risk mitigation strategies, and internal risk management and governance practices for GPAI model providers, with the aim of developing a code of practice for trustworthy GPAI models with systemic risks.

5.1 Defining Systemic Risk in GPAI Models

According to Article 51 of the AI Act, a GPAI model is classified as posing a systemic risk if either:

a) The GPAI model demonstrates high impact capabilities, as evaluated by state-of-the-art methodologies and tools, or

b) The European Commission determines that the GPAI model has comparable high impact capabilities.

The AI Act outlines multiple criteria to assess whether a model has high impact capabilities:

1. Technical evaluation:
   - using tools and benchmarks to assess the model's capabilities
   - using a computational threshold to determine if the training process required more than 10^25 floating point operations (FLOPs).
2. The commission assessment (as per Annex XIII) outlines the following criteria for consideration:
   -the number of model parameters
   -the quality and size of the training dataset
   -the computational resources used for training
   -input and output modalities
   -benchmarks and capability evaluations
   -market impact and reach (for example, if the model is made available to at least 10,000 registered EU-based business users)
   -total number of users

Understanding High-Impact Capabilities

Article 3(64) of the AI Act defines 'high impact capabilities' as those equal or exceeding the recognized capabilities of the most sophisticated GPAI models. This definition implies a flexible benchmark which will affect our understanding of what constitutes 'high impact' capabilities in accordance with future model developments. It also implies a benchmark that will be constantly moving because no reference moment is given to determine what exactly is the 'most sophisticated GPAI model'.

Relationship Between Model Capabilities and Systemic Risk

Art. 3(65) 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 the society as a whole, that can be propagated at scale across the value chain.

The AI Act establishes a link between advanced GPAI model capabilities and the likelihood of a GPAI model presenting systemic risk. Recital 110 states that systemic risks increase in tandem with increases in a GPAI model's capabilities and reach. This characterization suggests that a model's systemic risk is largely a function of its ability (due to its capabilities) to have either a broad scope or a significant magnitude of impact.

Factors Influencing Systemic Risk

In addition, recital 110 outlines factors that influence the emergence of systemic risks. These include but are not limited to conditions of misuse, model reliability, model fairness and model security, the level of autonomy of the model, its access to tools, novel or combined modalities, release and distribution strategies, the potential to remove guardrails, as well as the ability for these effects to propagate across the value chain.

Whilst the AI Act's approach to defining and assessing systemic risks stemming from GPAI models is multifaceted, considering technical capabilities, market factors, and health, safety, fundamental rights and societal impacts, it also introduces ambiguities and overly basic correlations. These issues could undermine the effectiveness of the regulation for GPAI model providers, particularly when applied to future GPAI models that may exhibit different characteristics, such as high generality or capacity, which are not adequately addressed by the current criteria. As such, this brief will suggest areas to broaden the scope of assessment to capture the potential evolving nature of GPAI models, and their associated developmental paradigms and better mitigate the risks associated with emerging GPAI models.

In sum, the AI Act states systemic risk should be understood to increase with model capabilities and model reach. In turn, high-impact capabilities, those which classify a GPAI model as having systemic risk, are assessed through a) Computational threshold (10^25 FLOP) b) Technical evaluations and methodologies c) Market impact.

5.1.1. Incorporating Level of Model Generality

The current AI Act framework for characterizing a GPAI model as having 'high impact capabilities' inadequately addresses the full scope of factors contributing to systemic risk. While the AI Act recognizes model capability and reach, it overlooks levels of model generality. The AI Act considers an AI model to have a ‘general purpose’ if it exhibits significant generality. What is considered ‘significant’ however remains unclear. Moreover, we argue that for the purpose of determining systemic risk, the level of generality (above the significance threshold) should be a factor of consideration.

Model generality refers to a GPAI model’s ability to perform across diverse tasks and environments (Hernández-Orallo, 2019:531). This concept is distinct from capability, which denotes a model’s proficiency in performing specific tasks or functions within a given environment (Burden & Hernández-Orallo, 2020:2). The degree of generality describes the model’s performance distribution over tasks of varying difficulty, including whether the model consistently performs across easy, moderate, and challenging tasks, or whether the complexity of the tasks changes the model’s performance (Hernández-Orallo et al., 2021:1).

Although a model’s capability in various areas can contribute to its generality, the two concepts remain distinct. Generality encompasses the model’s broader potential to adapt and excel across a wide variety of tasks and environments, rather than just performing several isolated tasks. High generality allows the model to apply learned knowledge to new and unseen tasks through methods like transfer learning, reflecting both depth and breadth (Rohlfs, 2024:14). In this way, GPAI models (operating with a significant level of generality) can exist on a spectrum varying in their capabilities and their generality.

Having established that generality is distinct from capability, we will now outline why increases in model generality (beyond the significance threshold) should be recognized as a factor that also increases systemic risk. A GPAI model with high generality can function proficiently across a broad range of tasks and environments, including those unforeseen by its developers (Triguero et al., 2023:5-6). This adaptability can lead to the emergence of unexpected capabilities, potentially transferring improved performance from one domain to another (ibid). Consequently, a highly general GPAI model can adapt to novel situations in ways that were neither explicitly programmed nor anticipated, increasing the likelihood of emergent capabilities and unforeseen risks.

The primary challenge with high generality lies in its unpredictability, especially in new environments. This unpredictability amplifies the potential for unintended consequences and complicates risk assessment, measurement, and mitigation. Testing and evaluating the model's capabilities, the ways in which the model might malfunction or produce undesirable outcomes across all possible scenarios becomes increasingly difficult.

Moreover, as generality increases in significance, the model's applicability could expand to a wider range of domains and tasks. This broad impact scope amplifies both potential negative and positive impacts across various sectors of society and stakeholders’ lives. From a technical safety perspective, models with high generality are also more challenging to control. For instance, as noted by researchers, traditional constraint methods like fine-tuning become less effective for highly general GPAI models, as training data can never capture all relevant situations.

As such, systemic risks associated with GPAI models are not solely a function of capability and reach, but also of level of generality. Higher generality can amplify the risks linked to increased capability and reach because a GPAI model that is not only highly capable and widely accessible but also highly general presents a more complex risk profile. This multidimensional nature of risk underscores the importance of considering level of generality alongside capability and reach in assessing and managing the systemic risks posed by GPAI.

Recommendation:
Systemic risk criteria should be expanded to include the level of model generality as a key factor in assessing systemic risk. This would better capture the full range of risks posed by highly adaptable and general GPAI models.

5.1.2. Including/Recognizing Unpredictable Systemic Risks

The AI Act's conception of risk is overly focused on predictable systemic risks, rather than recognizing the unpredictable systemic risks inherent to the nature of GPAI models. We propose broadening the scope of systemic risks beyond predictable capabilities to include unpredictable emergent capabilities and capacities.

Recital 111 states systemic risks result from "high-impact capabilities, evaluated on the basis of appropriate technical tools and methodologies, or significant impact on the internal market due to its reach". This approach relies on a model's demonstrated capabilities as a determinant of systemic risk. As previously outlined, a model's capabilities are its demonstrated abilities to perform specific tasks or sets of tasks based on current performance. As a result, the notion of existing capabilities is inherently predictive. However, by only considering systemic risks stemming from recognized and validated capabilities, we omit a field of potential systemic risks: those arising from unexpected impacts of GPAIs, particularly through emergent capabilities. Emergent capabilities are significant because they highlight the potential regulatory gap between a model's existing, demonstrated capabilities and its capacity for unexpected advancements. Indeed, emergent capabilities represent a class of unknown and unpredictable abilities.

In this way, we propose adding ‘potential (or capacity) for emergent capabilities’ as a subsection under extant capabilities in the list of systemic risk sources. Emergent capabilities are those not explicitly programmed by GPAI model providers, they can be capabilities that show unexpected or sudden performance improvements and can arise without specific training (ibid).

Recognizing a GPAI model's capacity for emergent capabilities, in addition to its extant capabilities, is important given that a significant portion of the risks associated with GPAI models stems from their inherent unpredictability (Boine & Rolnick, 2023:35). Indeed, unpredictability in terms of emergent capabilities is especially germane to the ways GPAI models function and are developed, particularly large language models (LLMs) (Wei et al., 2022) GPAI models are unpredictable because they are frequently retrained and updated, which increase the likelihood of unexpected emergent capabilities (Boine & Rolnick, 2023:35). Moreover, the outputs of LLMs are highly dependent on the inputs and specific contexts in which they function. This makes it impossible to test every conceivable input and context, complicating the identification and anticipation of all potential capabilities. As a result, a GPAI model that did not exhibit certain capabilities during testing might display them in different settings or after post-training enhancements (Bengio et al., 2024:3). For example, the emergence of compositional capabilities (a capability being composed of other capabilities) illustrates how new and unforeseen abilities can arise (Anwar et al., 2024:25).

In addition to the issue of a model's potential capacity for emergent capabilities, approaches on measuring and assessing demonstrable capabilities are flawed for several reasons. Firstly, Anwar et al. pointed out how researchers often make claims about the presence/absence of a capability based on whether the model is able to carry out tasks that supposedly require that capability (2024:16). This approach may miss capabilities that were not specifically tested for. Secondly, capabilities may manifest differently in various contexts or deployment scenarios, making it challenging to comprehensively assess a model's full range of capabilities.

In sum, we argue that the AI Act's approach of focusing predictable capabilities as a source of systemic risks is insufficient. Rather, GPAI model providers should additionally face code of practice requirements vis a vis a) a GPAI model's capacity for emergent capabilities and b)the possibility of undetected capabilities due to limitations in evaluation methods and framings.

Recommendation:
The Codes of Practice should demand assessment of not only a model's demonstrated capabilities but also its capacity for growth and adaptation, its capacity for emergent capabilities. This dual focus would provide a more comprehensive understanding of potential systemic risks.

5.1.3. GPAI Model Designation of High-Impact Capability

As previously outlined, a GPAI model is classified as a systemic risk if it has high-impact capabilities. There are three main avenues for the designation of high impact capabilities for a GPAI model: evaluation through methodologies and tools, assessment by the European Commission, and default designation based on a computational threshold at the level or exceeding 10^25 flops of compute for training.

The reliance on existing evaluation methodologies and tools to assess a model's capabilities is problematic due to the lack of sophisticated technical tools and methodologies to accurately assess advanced AI capabilities (Bengio et al., 2024:1) Additionally, there is the challenge of predicting and measuring emergent capabilities that may not be apparent during initial evaluations. In turn, The European Commission's assessment process, while potentially more nuanced, is only available to those models which show different capabilities to contemporary state-of-the-art GPAI models, and those models which have used less than 10^25 flops of compute for training.

As a result, we see that the flops threshold is likely to be the primary determinant for most models falling under the designation of having high impact capabilities. However, this approach is problematic because compute is an unreliable metric for discerning high-impact capabilities, and because reliance on computational power as a primary indicator of high-impact capabilities is a potential false correlation.

While previous advances along the LLM GPAI model paradigm demonstrated that increases in compute and data led to more performant models, this relationship is not guaranteed to hold indefinitely. Indeed, current scaling laws have begun to show diminishing returns whereby beyond certain thresholds, increasing compute yields marginal improvements in model capabilities (Anwar et al., 2024:22).

This has led many researchers to express doubts on the efficacy of scaling (Azhar, 2024). For instance, it remains unclear to what extent language models can acquire advanced reasoning and abstraction capabilities merely through increased scale (Anwar et al., 2024:24). Some studies suggest that the limitations of current large language models are unlikely to be resolved by scaling alone (Anwar et al., 2024:25).

In turn, high-impact capabilities often involve factors that go beyond raw computational power, such as novel architectures, improved training techniques, or algorithmic breakthroughs in model design. While easily measurable, the metric of compute doesn't capture the multifaceted nature of AI advancement and may lead to oversimplified assessments of model capacities. Indeed, alternative paths to capability improvements have been shown to arise through phenomena such as Grokking (where a model can improve capabilities while maintaining fixed compute and data resources) and in-context learning (Huang et al., 2024).

Although recital 111 outlines the provision for changing the threshold in lieu of further technical advancements, we argue that given these considerations, it appears misleading to rely on compute flops as a key determinant of whether a GPAI model possesses high-impact capabilities. This concern is salient given that the focus on compute-intensive models may overlook alternative AI paradigms that could achieve high-impact capabilities without relying heavily on computational resources. Sarah Hooker's concept of the "hardware lottery" suggests that future research directions may shift towards paradigms that bypass current scaling constraints (Hooker, 2021).

Recommendation:
Establish a flexible framework that can adapt to emerging AI paradigms and GPAI model advancements for systemic risk designation.

5.2 Systemic Risk Taxonomy (in response to Q1.10)

Having examined, interrogated, and identified issues in the AI Act's framework for characterizing GPAI models with high-impact capabilities and its narrow consideration of factors contributing to systemic risk, this paragraph will now offer several proposals for a systemic risk taxonomy in response to the EU's consultation questionnaire. Note that ALLAI does not intend to provide a limited list of proposals, but merely an initial one that can be supplemented in the future.

5.2.1. The Displacement of Humans in the Workplace and Deterioration of Skills

One systemic risk with foreseeable negative impacts on "society as a whole" and the fundamental "right to work" is the displacement of humans in the workplace. Multimodal GPAI models capable of generating media and content could potentially replace creative professionals in industries like entertainment and the arts, but also knowledge professionals in advertising, consultancy, and administrative professions. Studies have already begun to document the negative impact of LLMs on certain job categories (Tiwari, 2023).

Beyond impacts on job displacement/loss, this risk extends to the obsolescence of skills in certain professions, diminished job quality due to increased automation, and the potential exacerbation of inequalities (Korinek & Stiglitz, 2019).

This systemic risk pertains to GPAI providers by virtue of their responsibility in terms of the pace at which these models are deployed in society, especially regarding societal preparedness. Providers may also need to consider their responsibilities when deploying GPAI models internally, potentially replacing existing workers. Moreover, increased reliance on AI to advance GPAI models could create systemic risks related to dependency and the growing opacity (black box) of these processes.

5.2.2 The Systemic Risk of GPAI Model Concentration and Homogeneity

Another potential systemic risk with implications for "public and economic security" and "society as a whole" is the concentration of use of ever fever models, leading to 'homogeneity' of outcomes and single points of failure that can spread risks to numerous downstream applications. The lack of model diversity stemming from a limited number of (or even a single) GPAI model(s) becoming more advanced than others could stifle innovation. The foundational nature of many large scale GPAI models exacerbates these risks, as they often share similar or identical components, leaving them vulnerable to correlated failures and external attacks. For example, several studies have shown that the same jailbreak attacks can be transferred across different LLMs (Anwar et al., 2024:36). Moreover, the economic and power imbalances resulting from such concentration could undermine procedural and distributive fairness, potentially homogenizing cultural outputs and perspectives (Bommasani et al., 2022:149-152).

This systemic risk pertains to GPAI providers by virtue of their responsibility of managing the rate of deployment / release of their GPAI models in society, ensuring there is a stream of market diversity and that a certain GPAI model doesn't become overly dominant in the market.

5.2.3. Systemic risk to the environment

The impact that particularly the training of ever larger models has on the environment, due to its use of resources such as electricity and water is well known.

5.2.4. Systemic risk to fair competition and GPAI developers becoming 'too big to fail'.

Because access to critical resources for current GPAI models (large datasets, CPU, expensive chips) is ever more limited to ever fewer financially strong actors, fair competition becomes vulnerable. Moreover, when GPAI models infiltrate our sectors more widely and deeply, their few providers will eventually become 'too big to fail', a problem that led to the 2008 financial crisis, when banks had become 'too big to fail'.

Recommendations:
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'

5.3 Sources of Systemic Risk

Further to our answers to Question 11 of the Consultation, we propose the following additional factors to be considered as sources of systemic risks in relation to GPAI models. Note that ALLAI does not intend to provide a limited list of additional factors, but merely an initial one that can be supplemented in the future.

5.3.1. Human Error in GPAI Model Development

Human error can be a potent source of systemic risk in GPAI models as these models are ultimately designed, coded, and maintained by fallible humans. Errors can arise at any stage of development, from model specification and design to coding and ongoing monitoring. For example, a developer might inadvertently mis specify the model by choosing an inappropriate algorithm, misunderstanding relationships between variables, or neglecting critical environmental factors in model design (Bengio et al., 2024:2). Similarly, undetected bugs, such as typos or logic errors in the code, can lead to incorrect model behavior or poor design choices, introducing significant risks that may manifest on a systemic scale (Steimers & Schneider, 2022).

5.3.2. Deception in GPAI Model Oversight

As GPAI models become more sophisticated, they may develop capabilities to deceive oversight. Research has shown that advanced models can produce false but compelling outputs, complicating oversight and making monitoring more challenging (Bengio et al., 2024:3). In turn LLM agents have shown deceptive behavior despite attempts to train the models to avoid such behavior (Anwar et al., 2024:11). The systemic risk here lies in the difficulty of ensuring that these models are trustworthy and that their behavior aligns with intended ethical standards.

5.3.3. Unpredictable/unexpected jumps in Capabilities

The use of AI systems and alternative models to advance GPAI model capabilities introduces systemic risks due to the potential for rapid and unpredictable jumps in capabilities ((Bengio et al., 2024:2). When AI is used to optimize or improve a model, it may introduce changes and capabilities that are not fully understood by human overseers, leading to unforeseen changes in the model's behavior. While this does not necessarily mean a loss of control, it does raise concerns about the unpredictability of AI-driven enhancements and their systemic implications.

5.3.4. Systemic Risks specific to LLMs: Data Leakage & Model Hallucinations

Systemic risks can also arise from data leakage due to model malfunctions, posing significant privacy risks. In scenarios where a GPAI model leaks private information from one user to another, the impact can be widespread, especially if the model has a high reach and affects numerous stakeholders. For example, current LLMs do not reliably prevent such leaks, even when mitigation strategies like prompt engineering and output filtering are employed (Anwar et al., 2024:68).

Model hallucinations can represent a significant source of systemic risk, especially when decisions are based on incorrect or misleading information generated by the model. These hallucinations can have serious implications for fundamental rights, public health, and safety, depending on whether the erroneous output affects an individual or is propagated across interconnected processes.

5.3.5. GPAI Model (Mis)Alignment

Misalignment between a model's behavior and human intent is a source of systemic risk, encompassing more than just loss of control (as recognized in the consultation). Alignment failures can occur when it is difficult to formalize developer intent into precise model specifications, leading to reward hacking or goal mis-generalization (Anwar et al., 2024-32-33). For example, a model may optimize for a specified reward in a way that is misaligned with the developer's true intentions, resulting in behavior that appears correct but is fundamentally flawed (ibid).

5.3.6. Contextual and Organizational Dynamics

Finally, systemic risks can also emerge from contextual and organizational dynamics, which can be overlooked in favor of trustworthy AI discussions. For example, competition between GPAI providers can lead to destabilizing dynamics, such as a race for limited resources like compute power and data, which could exacerbate systemic risks.

Recommendations:
Add the following sources of systemic risk:

- Human error in GPAI Model development
- Deception in GPAI Model oversight
- Unpredictable/unexpected jumps in capabilities
- Specific to LLMs: Data leakage & model hallucinations
- GPAI Model (mis)Alignment
- Contextual and organizational dynamics

5.4 Systemic Risk Identification Assessment Measures

Further to our answers to Question 13 of the Consultation, we propose the following additional measures for risk assessment to be considered in relation to GPAI models. Note that ALLAI does not intend to provide a limited list of additional measures, but merely an initial one that can be supplemented in the future.

5.4.1 Provision for Flexible Risk Thresholds and Risk Tolerance

While we underscore the importance of determining and establishing appropriate risk thresholds and tolerance levels for GPAI model development, it is equally important to ensure these thresholds remain flexible and dynamic in response to new advancements and evidence. Given the pace of GPAI model development, fixed risk thresholds may become outdated which can compromise the effectiveness of risk management practices. As new capabilities and application domains emerge, there should be a continuous reassessment and adjustment of risk thresholds to reflect the latest developments. In addition, as new evaluation methods and techniques are being developed, risk thresholds should also be adjusted to reflect the latest knowledge and evidence.

As such, alongside determining risk thresholds, tolerance levels, and quantifying risk severity and probability, providers should be required to regularly evaluate and update their risk frameworks to ensure they stay relevant and effective. Moreover, providers should notify the appropriate (AI regulatory authorities) when new knowledge or evidence of capabilities arise that may impact the established risk thresholds, or if the thresholds have become inadequate.

5.4.2 Cumulative Risk Assessment and Forecasting

We take issue with the current delineation of systemic risk, arguing the boundary between localized and systemic risk is often blurred. Indeed, it seems possible that systemic risks can emerge even when GPAI models have limited reach or seemingly limited capabilities.

We propose that providers should consider the cumulative and aggregate nature of GPAI model-related risks. In this view, small-scale effects, seemingly isolated risks or individual-level issues, which are not immediately considered systemic risks, can aggregate overtime to create larger systemic risks.

As a result, we propose that providers should conduct GPAI cumulative risk assessments that account for the aggregation of minor risks that may compound over time, cascading effects where one amplifies other risks, and synergistic interactions between different risk categories. For example, providers could implement forecasting methodologies to project potential cumulative risks over extended time frames by considering various scenarios of GPAI model deployment and development.

5.4.3 Long-Term Systemic Risk Assessment

GPAI models have the potential to cause systemic risks that may not immediately manifest but could have profound long-term implications for society. These risks could develop gradually, making them challenging to detect and address in early stages. As a result, we propose that providers should extend risk assessment frameworks to incorporate longer time horizons to consider potential negative effects that may unfold over years. For example, these assessments could include the forecasting of gradual shifts in employment structures and labor markets, long-term effects on educational systems, potential shifts in social norms and dynamics, and impacts on democratic processes over time.

Recommendations:
Add the following risk assessment measures:

- Flexible risk thresholds and risk tolerance
- Cumulative risk assessment and foresight
- Long-term continuous systemic risk assessment

5.5 Systemic Risk Evaluation Practices

Further to our answers to Question 15 of the Consultation We propose these additional GPAI model evaluation practices to to effectively evaluate systemic risks along the entire model lifecycle. Note that ALLAI does not intend to provide a limited list of additional measures, but merely an initial one that can be supplemented in the future.

5.5.1 Continued Model Trustworthiness Evaluation

While evaluating the capabilities of a GPAI model is necessary, providers should also conduct explicit trustworthiness evaluations. Although methods to evaluate whether a model is performing and behaving as intended are still developing, several promising approaches are emerging. For example, red teaming involves simulating adversarial scenarios to challenge the model. This process evaluates whether the model remains aligned and behaves safely. Currently however, red teaming approaches predominantly look at model behavior from a safety perspective. We strongly recommend however to include all elements of AI trustworthiness (lawfulness, ethical alignment and robustness) are included prominently.

5.5.2 Model Generality Level Evaluations

Evaluating the level of generality of GPAI models is important due to the potential for unexpected and emergent capabilities that can arise from their broad applicability and generalization ability. However, many current GPAI model evaluations, particularly for LLMs, remain domain specific. This limitation stems from the logistical challenges of assessing LLMs across the vast array of possible domains and tasks. Consequently, researchers are advocating for alternative methods to more effectively evaluate generality and cross-domain generalization.

To address these challenges and mitigate systemic risks associated with the level of generality of GPAI models, several promising state-of-the-art technical evaluation methods should be integrated into the evaluation process. One such method is Skill-Mix, a procedural evaluation technique designed to assess the compositional generalization abilities of GPAI models like LLMs (Yu et al., 2023). This technique evaluates the model's capacity to combine previously learned skills and apply them to new, unseen tasks, thereby providing insights into the model's adaptability and generalization across different contexts (ibid:[ ]).

Another important approach is mechanistic investigations, which focus on identifying and analyzing capabilities that are reused across various tasks. These investigations identify the underlying mechanisms that contribute to the model's general-purpose behaviors, such as in-context learning, offering understanding of how the model generalizes its capabilities across different domains.

Recommendations:

Add the following risk evaluation measures:

- Continued model trustworthiness evaluation
- Model generality level evaluation

∞

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