Considerations regarding a tiered approach for foundation models and general purpose AI

Summary

- **We caution against a multi-tier framework for regulating foundation models, GPAI and generative AI with too many overlapping layers of stringent obligations**, with not clearly defined concepts to distinguish between the tiers and thresholds that given a very nascent field of research might not be the best proxies for measuring risk stemming from AI. Hasty and substantive changes to the AI Act, without proper assessment, evidence, and discussion risk falling short of desired objectives, and might lead to unintended consequences for providers, deployers and users.
- We recommend the envisaged tier for **'general purpose systems at scale' to be rejected**. The proposal for GPAI at scale creates substantial overlaps and legal uncertainty with the remaining parts of the AI Act, in particular in conjunction with specific rules for 'very capable foundation models'.
- **The regulation of GPAI** should be confined to systems deployed (or serving as components) in high-risk applications. Separately, the European Commission should establish a code of practice in collaboration with industry and AI experts for those developing the most advanced AI models focused on Foundation Models.
- **Threshold criteria** relating to compute power (FLOPs), training data, number of users or training data are unsuitable to determine risk as they neglect the actual outputs or risk of a model. Performance based benchmark tests and evaluations are more appropriate as they take account of safety measures and establish the closest approximation of risk. Think of crash tests for cars instead of assessing details of their production. Performance evaluations and benchmarks are not commonly established or defined and require close collaboration with experts, providers and regulators.
- Foundation models benefit from **risk assessment and mitigation**. This could include – for example – internal red-teaming, but rather than prescribing the precise method, Requirements should therefore take account of available expertise, the absence of recognized standards, and crucially, provide ample safeguards that ensure testing is actually workable.
- Too descriptive or invasive **requirements**, such as external testing before and after marketing and regardless of identified risk will result in an unworkable framework. Data, including trade secrets, proprietary and security relevant data ought to be protected from disproportionate disclosure, and access to proprietary systems needs to remain a measure of last resort.
- **Transparency** should aim to build trust and understanding of AI. Before mandating transparency and specific technical specifications (labels, watermarks, detection or provenance) it should be ensured that these means contribute substantially to building

trust and understanding, to avoid technical requirements becoming an end of their own. As technical solutions are nascent and experimental, rules should remain voluntary.
- **Governance** should aim to establish harmonised, comparable, consistent and effective outcomes across the EU. Centralised enforcement should, above all, contribute to better outcomes instead of adding complexity, cost or legal uncertainty. The responsibilities and powers across all regulatory authorities should be consistent and in line with the EU market surveillance framework and include instances to include expertise from the affected providers and deployers, as well as experts and international stakeholders. Such collaborative approaches ensure state-of-the art developments being reflected and broadly accessible.

To address the open **questions and challenges around a tiered approach**, in particular to:

- Ensure a proper understanding of the risk the AI Act is trying to address;
- Based on that risk, define the appropriate metrics that correlate risk with output and performance;
- Uphold the legal basis of the AI Act, a uniform protection of fundamental rights, and a grounding in the risk-based approach;
- Model legislation that captures the nature of quickly evolving technology, research and international consensus;
- Propose a workable enforcement of the law;

We believe that the **following approach** may describe a way forward:

If an additional tier is required, this should focus on performance (i.e. the comprehensive capabilities) of foundation models assessed through output-evaluations, not compute, user number or training data. To allow for a future proof, internationally aligned and evidence based approach, the AI Act should specify the objective of addressing new risks that relate to new capabilities of foundation models; and the need to develop proportionate mitigation measures. The details of assessing risks, defining evaluations to determine which models are in scope, and appropriate mitigation measures should be delegated to voluntary codes or similar fora that allow to develop fit for purpose metrics.

More in detail:

Introduction

The tiered approach refers to the idea to align the respective positions of the Council and the European Parliament regarding the regulation of foundation models and general purpose AI systems. This involves a common level of requirements across all foundation models and

specific requirements applicable for “very capable foundation models” and “general purpose AI systems at scale”. This note discusses **principles** around a tiered approach, **challenges** of approximating risk with size thresholds, challenges with **proposed requirements**, **more suitable parameters** to approximate risk, and **a possible way forward** to regulating general purpose systems in a fast evolving environment.

Principles:

- **We caution against a multi-tier framework for regulating foundation models, GPAI and generative AI with too many overlapping layers of stringent obligations**, with not clearly defined concepts to distinguish between the tiers and thresholds that given a very nascent field of research might not be the best proxies for measuring risk stemming from AI. Hasty and substantive changes to the AI Act, without proper assessment, evidence, and discussion risk falling short of desired objectives, and might lead to unintended consequences for providers, deployers and users
- **As a general remark,** we remain convinced that views reflected by the Council that a purely risk-based approach based on systems and their use case is a better fit for the legal structure of the AI Act and product safety. Safety, quality and potential impact on fundamental rights depend on the specific use of the GPAI in an application. To stick to the risk based approach, **the AI Act should endeavour be limited to regulating GPAI only when deployed in high-risk uses** and allow exemptions if only used for low risk first or third party applications.
- **A tiered approach to the underlying technology moves away from the AIA’s risk-based approach,** undermining the careful balance of innovation and safety that was the original intent of the legislation. The size or popularity of a model or system does not predict its level of risk. This is a radical departure from the original approach of the legislation.
- Defining tiers via **thresholds that rely on size, user number, compute or data** will fail to identify risk and create loopholes and inconsistencies.
- Moreover, without a clear definition of the perceived risks with foundation model and GPAI, the tiered-approach suffers from a fundamental flaw, namely that without a clear objective for regulating a particular product it is impossible to draft appropriate and proportionate requirements.
- **This approach is out of step with international co-regulatory approaches** which seek to promote innovation, recognize the fast-moving pace of research and development and ensure societal values and fundamental rights are protected. The AI Act should be consistent and compatible with international efforts and avoid duplication.

Overlapping categories of models that are poorly defined will create uncertainty

- A tiered approach that addresses two separate sets of categories for GPAI systems and foundation models is confusing, out of step with industry developments and terminology and will create significant legal uncertainty. By creating multiple regimes in parallel creates overlap in requirements, hence adding further confusion to an already complex regulation and will provide no legal clarity to developers or deployers of these systems in the EU.
- In particular, it is **unclear what distinguishes a non-very-capable foundation model from a GPAI**. This will lead to unnecessary confusion for the industry.
- Moreover, **no specific risk has been identified with regards to GPAI (or GPAI "at scale")** which, in turn, makes the definition and regulation of these systems highly arbitrary. For example, if the legislator is concerned by generative AI in particular, then it should clearly state so and then a meaningful discussion can be had on particular risks and whether those are already addressed by other parts of the AI Act or existing EU legislation. Instead, the proposal targets an open-ended category of GPAI for special regulatory rules, despite the fact that GPAI -- or even GPAI "at scale" -- covers disparate types of AI with fundamentally different capabilities and risk profiles .
- Regulators must clearly define the categories of systems, products and tools that will be subject to regulation or they will introduce so much ambiguity and uncertainty that it will become impossible to develop or deploy models in the EU.

How to fix: We recommend the envisaged tier for 'general purpose systems at scale' to be rejected. The proposal for GPAI at scale creates substantial overlaps and legal uncertainty with the remaining parts of the AI Act, in particular in conjunction with specific rules for 'very capable foundation models'. The regulation of GPAI should be confined to systems deployed (or serving as components) in high-risk applications. Separately, the European Commission should establish a code of practice in collaboration with industry and AI experts for those developing the most advanced AI models focused on Foundation Models.

Size-based thresholds to define categories of models is a flawed methodology; performance-based benchmarks would be a more appropriate approach:

- **Compute is not a good proxy for identifying "very capable foundation models"**: There is no direct link between the amount of compute used for training (FLOPs) and the potential risk stemming from a foundation model. Although there is -- as of now -- an association between model scale and capabilities, that is only true if the basic model architecture, training algorithm, and dataset are all held constant. Changing one of those components can result in better model performance with fewer FLOPs.

Moreover, innovations in gradient descent algorithms over the past few years have made it possible to maintain performance with fewer FLOPs. **Even when thresholds are regularly updated, they risk overlooking models that actually present a risk.** Relying on any threshold of compute alone, will risk that less powerful but potentially unsafe models remain outside the scope of the AI Act.

- FLOPs describe the computational power that went into training a model, but compute requirements for training do not reliably predict the risk level of a model. Assume two models are trained using the same amount of FLOPs. One model undergoes careful data governance, using tools and datasets to identify and mitigate bias; examining data for accuracy, completeness, labels, redundancies, etc; and is continuously tested for safety and being evaluated after deployment. The other model is not submitted to the same level of scrutiny and safety testing. While the FLOPs are the same, the safety of the first model would hardly be comparable to the second model, which didn't undergo the same rigorous testing. **Relying on any threshold of compute alone, will risk that less powerful but potentially unsafe models remain outside the scope of the AI Act.**

> *A common misconception in relation to FLOPs or model size is that they linearly reflect performance, which however is wrong. FLOPs and model size only reflect performance as long as all other parameters, including model architecture, training algorithm, dataset and model weights are constant. In reality however, all these components constantly change which makes the validity of FLOPs or model size noisy and imprecise proxies. This has been shown by recent innovations in gradient descent algorithms over the past years (e.g. LoRA) - these have made it possible to maintain performance with fewer FLOPs. Smaller models like LLaMA or Mistral 7B have shown surprising performance with smaller size. A different example are models that include 'distilled' behaviour of larger models, instead of training from standards datasets. Risk profiles of such models vary substantially based on if and which additional safety protocols are deployed at the output level.*

- **The number of users is a poor signifier and unstable measure of potential risk:** It seems similarly futile to define the impact of a model or system based on the number of downstream business or consumer users - given that these numbers do not allow for conclusions about risk, vary over time and are known only after a model has been deployed. While the probability of harm from known risks from a given system might increase with the scale of deployment of that model, it is the a priori presence of harmful capabilities that drive risks, not scale alone. A model or GPAI used by a small number of users might still have large scale consequences if those users are decision-makers in critical sectors and the application presents specific risks. The number of users does not impact how individuals use a system; users with malicious intent are likely to seek out applications that allow for unintended or harmful use

regardless of mass adoption. Conversely, millions of users could use a model for trivial tasks with minimal societal impact. In the case of models, it is unclear how users may be counted at all since the same model is likely deployed across a multitude of different applications downstream and by third-party deployers.

- **Training data does not indicate the potential risks created by a model:** Possible secondary measures such as the amount of data used and the number of high-risk applications a model is deployed in are imprecise measures, too. Research shows that the amount of data used in LLMs varies and, as explained above, even models with smaller training sets can bear risk. Conversely, large language models with large training datasets can be fine-tuned in particular directions, including towards higher risk uses, which makes the training data an even less useful object of analysis. Especially if datasets are incomplete or overly biased as a result of the smaller datasets. Using the amount of high-risk applications a model is deployed in risks having a disproportionate effect on model providers which would be subject to two different regimes under the AI Act, enforced by two different regulators.

Indeed, If the intention is to define risk, i.e. dangerous capabilities of a model - these are likely entirely disconnected from size, as 'small' models built on limited compute might cause substantial risks that will not be reflected in such a threshold

Performance-based benchmarks a are a more effective means of categorization

In an effort to understand the potential impact of foundation models or GPAIs beyond their specific use better, one could focus on the **nature and risks of this type of product,** rather than any external elements. So what distinguishes 'very capable' foundation models from other types of AI? It is primarily their capability to perform significantly better than other AI systems and across a wide range of tasks, i.e. they possess **new capabilities** that could present **new risks**. It is important that capability is defined through the risk of outputs, not the size/amount of inputs. It's important to reiterate that model size or compute is not a suitable proxy for risk emerging from such models, which depends on many other factors as outlined above.

- **Benchmarks and evaluations** are more suitable to test the risk of a given model. Benchmarking a model takes account of the effective risks, including risks that are not yet known. Benchmarking also takes account of risk mitigation and safety features that are deployed in a model.

Benchmarks and evaluations in this field are still very nascent and likely to substantially evolve in the near future. Hence any thresholds to determine performance or capabilities will need continuous assessment and updating as these capabilities will continuously evolve. Any provision would need to allow us to assess risk dynamically, besides the possibility to update

thresholds.

Given that the technology behind foundation models is still rapidly developing, the AIA could become obsolete very quickly if it were to define the capability indicators in the law, for example, if the legislator would have set – through implementing acts concrete tools and methodologies to predict and measure the capabilities – it seems certain that those would have been outdated already today. While conceptually, performance and capability could be clarified in the text, codes of conduct are most suitable to further define and update methodology, thresholds and technical details.

- **Capability or performance based on benchmarks are a more suitable measure to define advanced models that require more scrutiny**. While not representative of risk or impact on fundamental rights, capabilities may give an indication about certain aspects linked to technological novelty and potential future risks of a foundation model. Referring to dynamic capability / performance indicators are less likely to cause unintended consequences compared to external indicators such as size or users.

It's important to note that while performance-based criteria are a better representation of model capabilities, **capabilities remain different from risk or impact on fundamental rights**. Less capable models might present increased risk if they are deployed in high-risk areas, if they are deployed towards nefarious goals, or released in certain modalities, which may allow for the removal of built-in safety filters.

Generally, any definition or threshold should be established through an expert process organised by an expert authority, be informed by evidence through due process and consultation, be based on emerging international standards and scientific research.

How to fix: Criteria relating to compute power (FLOPs), training data, number of users or training data are unsuitable to determine risk as they neglect the actual outputs or risk of a model. Performance based benchmark tests and evaluations are more appropriate as they take account of safety measures and establish the closest approximation of risk. Think of crash tests for cars instead of assessing details of their production. Performance evaluations and benchmarks are not commonly established or defined and require close collaboration with experts, providers and regulators.

Requirements should remain balanced with the potential risks of highly capable models

Ensuring the safety and security of AI models is an important goal of the AI Act. But the field is still new and consensus standards and best practices do not currently exist to guide policy

approaches – many of these remain open research questions, and a level of flexibility is needed to identify the best ones. We recommend the EU focus first on driving the development of these standards and best practices through inclusive, multi-stakeholder fora .

- **Transparency on risk management practices:** Providers already prepare technical documentation which can include information around risk governance, hardening measures, testing methodologies, and standards and benchmarks which have been adhered to. However, this transparency must be balanced against the risk of disclosure of sensitive security information and intellectual property, which can place EU citizens and organisations at risk. Disclosure should require information about the policies and practices applied, but not the specific risks themselves.

- **Red teaming best practices are evolving, and mandates for universal red-teaming by external parties are disproportionate:** Red teaming is an important subset of risk assessment and mitigation practices. We generally welcome such efforts as a way to test our systems and ensure a high level of safety. However, given the sensitivity of providing access to models, in particular ahead of broader model release, we urge caution about several aspects. As in many other fields of AI, red-teaming is evolving and there are no settled standards across the industry, including on the level of access to be provided to testers; which categories of vulnerabilities should be tested for; or how to responsibly disclose identified risks. There are also substantial concerns about the availability of sufficiently qualified personnel to conduct such evaluations and confidentiality might raise concerns about individual testers gaining insights into multiple, competing companies' proprietary information.

- **Overly descriptive disclosure requirements threaten quality of output, trade secrets, and security of citizens:** We urge caution regarding transparency and disclosure. Widely disclosing vulnerabilities identified through red team exercises could place European citizens and organisations at risk. Furthermore, results of red-team evaluations are core elements of proprietary data and therefore an overbroad obligation to involve external testers would conflict with good standards on trade secrets, proportionality and confidentiality. Again, disclosure should not become an end of itself but focus on relevant information. For this reason in lieu of requiring the sharing of red team results, we recommend requiring that providers have an approach to red teaming in place, and require disclosure of red team methodology, processes, procedures.

Compliance controls should similarly adhere to proportionality, due process and confidentiality standards to achieve a balanced outcome. The AI Act is principled in product safety which foresees self-assessments but allows regulators to ensure and check if such assessments live up to the letter of the law. In absence of a concrete risk, as this is the case for foundation

models of any capability, this concept should remain in place.

How to fix: Risk assessment and mitigation are important pillars to safety. This could include – for example – internal red-teaming, but rather than prescribing the precise method, Requirements should therefore take account of available expertise, the absence of recognized standards, and crucially, provide ample safeguards that ensure testing is actually workable. Too descriptive or invasive requirements, such as external testing before and after marketing and regardless of identified risk will result in an unworkable framework. Data, including trade secrets, proprietary and security relevant data ought to be protected from disproportionate disclosure, and access to proprietary systems needs to remain a measure of last resort.

Transparency should remain proportionate and protect trade secrets:

We strongly support development and deployment of mechanisms that enable users to understand if content is AI-generated, including robust provenance, watermarking, or both, however we have to be cognizant of the effectiveness and the technical feasibility of such proposals. In terms of effectiveness it is crucial to note that labels, watermarks or meta-data are a tool and can be deployed equally for legitimate or illegitimate ends. While trust should be the objective, technical solutions should not become an end of their own. We would argue that more work is needed to coalesce around robust, scalable, useful solutions before any method (e.g. watermarks, standards, interoperability, etc) is required by law.

Secondly, technologies for attribution of AI-generated content are nascent and currently best suited to limited modalities. While we are conducting experiments ourselves and have updated our content policies to account for the increasing prevalence of synthetic content, there is much to learn still on best practices and appropriate solutions across the ecosystem. Until there is clarity on appropriate solutions for all types of providers and deployers we recommend these efforts to remain voluntary and driven by industry standards.

How to fix: Transparency should aim to build trust and understanding of AI. Before mandating transparency and specific technical specifications (labels, watermarks, detection or provenance) it should be ensured that these means contribute substantially to building trust and understanding, to avoid technical requirements becoming an end of their own. As technical solutions are nascent and experimental, rules should remain voluntary.

A Harmonized and Consistent Approach to Governance of the AI Act is Required:

We welcome efforts to ensure uniform and highly qualified enforcement and regulatory oversight. This increases the effectiveness of legislation and ensures it is enforced proportionately, upholding due process and procedural standards. In this sense we welcome efforts around an AI Office that ensures consistent and highly qualified compliance across the EU, as well as efforts for the Office to be an interlocutor for providers that are subject to requirements and inviting expertise from academia and other stakeholders; also efforts to align European efforts internationally can improve the outcome of regulatory action and ensure consistency and compatibility with international law, and comparable efforts to govern AI.

However we urge caution to apply different regulatory instances across the AI Act. By having different tiers which partially are enforced by the Office and at the same time when the remaining, original, logic of the AI Act is based on enforcement by at least 27 competent authorities is increasing fragmentation, creating additional risk of inconsistent application and increasing the cost of enforcement for Member States and the Commission and undermining the Digital Single Market objectives.

More generally, we understand processes about compliance, enforcement and accountability as carefully established and balanced outcomes that both ensure effective compliance with legislation as well as proportionate and workable solutions for companies. As well established, for example in the DSA, regulators should have means to ensure compliance, however the requirement to test products ahead of their launch or continuously after their marketing seems disproportionate and raises serious concerns with regard to confidentiality, proprietary knowledge, and the effectiveness of such requirements.

How to fix: Governance should aim to establish harmonised, comparable, consistent and effective outcomes across the EU. Centralised enforcement should, above all, contribute to better outcomes instead of adding complexity, cost or legal uncertainty. The responsibilities and powers across all regulatory authorities should be consistent and in line with the EU market surveillance framework and include instances to include expertise from the affected providers and deployers, as well as experts and international stakeholders. Such collaborative approaches ensure state-of-the art developments being reflected and broadly accessible.

Focus on foundation models through a co-regulatory process for a future-proof regime:

Given the highly dynamic nature of AI development, uncertainties around risks and a fast moving international dimension, ensuring a future proof and sufficiently adaptable framework around the AI Act is crucial. Maintaining the AI Act's risk-based approach will enable the EU to

harness the opportunities of AI while mitigating the risks.

Any introduction of new tiers that move away from the risk-based approach should focus on one clearly defined set of models. The introduction of a set of rules for Foundation Models and GPAI, without clearly defining either category, will lead to an unworkable regime. We propose focusing on Foundation Models, to be further defined through a process described below, that would provide much needed certainty for the ecosystem.

The AI Act should remain open to certain non use-specific risks, even if those are not clearly specified or known today, and that such risks are best reflected through rigorous benchmarking and evaluations of model outputs, regardless of their size, compute power, user numbers, etc. Benchmarks should aim to identify *models that possess materially new capabilities that could present new safety risks compared to state-of-the-art foundation models*.

For such cases, and to be able to harness advances from the broader AI community, the European Commission should establish a code of practice in collaboration with industry and AI experts for those developing the most advanced AI models. This will enable leading experts to establish a clear understanding of risk and codify how this can be identified and mitigated. Such a code would enable the agreement on key principles to govern these very capable models including:

In summary, to address the open questions and challenges around a tiered approach:

- Ensure a proper understanding of the risk the AI Act is trying to address;
- Based on that risk, define the appropriate metrics that correlate risk with output and performance;
- Uphold the legal basis of the AI Act, a uniform protection of fundamental rights, and a grounding in the risk-based approach;
- Model legislation that captures the nature of quickly evolving technology, research and international consensus;
- Propose a workable enforcement of the law;

We believe that the following approach may describe a way forward:
If an additional tier is required, this should focus on performance (i.e. the comprehensive capabilities) of foundation models assessed through output-evaluations, not compute, user number or training data. To allow for a future proof, internationally aligned and evidence based approach, **the AI Act should specify the objective of addressing new risks that relate to new capabilities of foundation models; and the need to develop proportionate mitigation measures. The details of assessing risks, defining evaluations to determine which models are in scope, and appropriate mitigation measures should be delegated to voluntary codes or similar fora that allow to develop fit for purpose metrics.**