First Draft General-Purpose AI Code of Practice

Opening statement by the Chairs and Vice-Chairs

As the Chairs and Vice-Chairs of the four Working Groups, we hereby present the first draft of the General-Purpose AI Code of Practice under the AI Act (the "Code"). Written feedback by the Code of Practice Plenary participants and observers is welcome by Thursday, 28 November, 12:00 CET through a form on the dedicated platform (Futurium).

This first draft of the Code addresses key considerations for providers of general-purpose AI models and for providers of general-purpose AI models with systemic risk, through four Working Groups working in close collaboration:

- Working Group 1: Transparency and copyright-related rules
- Working Group 2: Risk identification and assessment for systemic risk
- Working Group 3: Technical risk mitigation for systemic risk
- Working Group 4: Governance risk mitigation for systemic risk

We present this first draft as a foundation for further refinement. **Following an iterative process of internal discussions within the Working Groups and additional external input from stakeholders, Measures may be added, removed, or modified**. We invite stakeholders to review the document and provide feedback to help shape the final version of the Code, which will play a crucial role in guiding the future of general-purpose AI model development and deployment.

We have included a high-level drafting plan that outlines our guiding principles and objectives for the Code. Although the first draft is light in detail, this approach aims to provide stakeholders with a clear sense of direction of the final Code's potential form and content, while we continue to engage in thorough deliberations regarding specific Sub-Measures and Key Performance Indicators (KPIs). To provide even more insight into our deliberation, we have added open questions to highlight some of the areas where we aim to make progress in future drafts. This also serves the purpose of guiding feedback and submissions to allow various stakeholders to continue to participate effectively.

The EU AI Act came into force on 1 August 2024, stating that the final version of the Code should be ready by 1 May 2025. The first draft presented here seeks to provide a "future-proof" Code that is also appropriate for the next generation of models that will be developed and released in 2025 and thereafter. Our work, which began in October 2024, has involved synthesising input from diverse stakeholders and engaging in iterative discussions.

In formulating this first draft, the Chairs and Vice-Chairs have been principally guided by the provisions in the AI Act as to matters within the scope of the Code. Accordingly, unless the context and definition contained within the Code indicates otherwise, the terms used in the Code refer to identical terms from the AI Act. This includes the instructions in Article 56(1) of the AI Act to take into account various existing international approaches. We have not included exhaustive references to provisions in the AI Act in this first draft, but expect to do so in future iterations.

This first draft of the Code is the result of a collaborative effort involving hundreds of participants from industry, academia, and civil society. We have also been informed by the evolving literature on AI Governance, international approaches, Union law codes of practice (such as the Code of Practice on Disinformation), and the expertise and experience of Working Group members.

Key features of the development process include:

- A multi-stakeholder consultation with nearly 430 submissions so far
- Four specialised working groups led by Chairs and Vice-Chairs selected for their expertise, experience, independence, and to ensure geographical and gender diversity
- Discussions and drafting sessions to be held between October 2024 and April 2025

Additional time for consultation and deliberation – both externally and internally – will be needed to refine and improve the current draft. As a group of independent Chairs and Vice-Chairs, we strive to make this process as transparent and accessible to stakeholders as possible, aiming to share our work and our thinking as early as possible while taking sufficient time to coordinate and discuss key questions within Working Groups. We count on your continued engaged collaboration and constructive criticism.

We welcome written feedback by the Code of Practice Plenary participants and observers welcome by Thursday, 28 November, 12:00 CET, through a form on the dedicated platform (Futurium).

Thank you for your support!

Nuria Oliver *Working Group 1* *Co-Chair*	**Alexander Peukert** *Working Group 1* *Co-Chair*	**Mattias Samwald** *Working Group 2* *Chair*	**Yoshua Bengio** *Working Group 3* *Chair*	**Marietje Schaake** *Working Group 4* *Chair*
Rishi Bommasani *Working Group 1* *Vice-Chair*	**Céline Castets-Renard** *Working Group 1* *Vice-Chair*	**Marta Ziosi** *Working Group 2* *Vice-Chair*	**Daniel Privitera** *Working Group 3* *Vice-Chair*	**Anka Reuel** *Working Group 4* *Vice-Chair*
		Alexander Zacherl *Working Group 2* *Vice-Chair*	**Nitarshan Rajkumar** *Working Group 3* *Vice-Chair*	**Markus Anderljung** *Working Group 4* *Vice-Chair*

Drafting plan and principles

At this stage, this first draft does not contain the level of granularity that will be included within the final adopted version of the Code, since: i) we strive towards broad agreement on the structure and principles of the Code; ii) there has been insufficient time to produce detailed proposals with the level of consideration in this first draft that such proposals would require; and iii) we will update the (draft) Code's details last to reflect latest developments on an ongoing basis. To illustrate our sense of direction, notwithstanding the generally high-level nature of this first draft, we include some glimpses of more concrete provisions under some commitments to show the form and detail that similar provisions may take in future drafts of the Code. The structure of the commitments in this Code follows a descending hierarchy of Measures, Sub-Measures and KPIs. Where any of these are missing, particularly KPIs, this is not a final decision but a consequence of the time constraints and high-level nature of this first draft. Moreover, this first draft does not yet contain a section on how the Code will be reviewed and updated – this will be incorporated in later iterations of the draft Code.

We also set out below some high-level principles we propose to follow when drafting the Code:

I. **Alignment with EU Principles and Values –** Measures, Sub-Measures and KPIs should be in line with general principles and values of the Union, as enshrined in EU law, including the Charter of Fundamental Rights of the European Union, the Treaty on European Union and Treaty on the Functioning of the European Union.

II. **Alignment with AI Act and International Approaches –** Measures, Sub-Measures and KPIs should contribute to an appropriate application of the AI Act. This includes taking into account international approaches (including standards or metrics developed by AI Safety Institutes, or standard-setting organisations), in accordance with Article 56(1) AI Act.

III. **Proportionality to Risks –** Measures, Sub-Measures, and KPIs should be proportionate to risks, meaning they should be (a) suitable to achieve the desired end, (b) necessary to achieve the desired end, and (c) should not impose a burden that is excessive in relation to the end sought to be achieved. Some concrete applications of proportionality include:

   a) Measures, Sub-Measures and KPIs should be more stringent for more significant risks or uncertain risks of severe harm. The Code can accomplish this by, for example, suggesting a multitude of KPIs for each Sub-Measure related to a severe risk, thereby requiring providers of general-purpose AI models to take action to mitigate that severe risk or to robustly demonstrate an extremely rare likelihood of severe risk eventuating. The Code might also tie risk-mitigating Sub-Measures to risk-assessment KPIs, such as using "if-then" requirements. For example, if a general-purpose AI model with systemic risk is assessed to have capability X, Y risk mitigations must be in place, guided by Z KPIs.

   b) Sub-Measures and KPIs should be specific. We accept that Measures may be written at a higher level of generality than Sub-Measures and the KPIs evidencing compliance with Sub-Measures. Nonetheless, general-purpose AI model providers should have as clear of an understanding as possible of how to satisfy Sub-Measures, as appropriate, evidenced by KPIs. Sub-Measures and KPIs should also be robust to circumvention or misspecification. The Code can accomplish this by, for example, avoiding the unnecessary use of proxy terms or metrics. The AI Office will monitor and review Sub-Measures and KPIs that may be susceptible to circumvention and other forms of misspecification.

c) Measures, Sub-Measures and KPIs should differentiate, where applicable, between different types of risks, distribution strategies, deployment contexts, and other factors that may influence the level of risk, and how risks may need to be assessed and mitigated. For example, Measures, Sub-Measures and KPIs related to assessing and mitigating systemic risks might need to differentiate between intentional and unintentional (including misalignment) risks, and might be more or less specific and stringent for some types of risks, distribution strategies (e.g. open-sourcing), and deployment contexts than for others.

IV. **Future-Proof** – Sub-Measures and KPIs should preserve the AI Office's ability to improve its assessment of compliance based on superior information. Moreover, the process for updating Sub-Measures and KPIs should assume that rapid technological change may require agile regulatory development and modification. Therefore, a balance should be struck between concrete requirements and flexibility to adapt and update rules as technology and industry develops. The Code can accomplish this by, for example, referencing dynamic sources of information that providers can be expected to monitor and consider themselves. Examples of such sources could include incident databases, consensus standards, risk registers, risk management frameworks, and AI Office guidance. The Code shall strive to enable Sub-Measures and KPIs to be more rapidly updated than usual. It may also be necessary to articulate types of models that would require a new set of Sub-Measures and KPIs, for example, models used in agentic AI systems.

V. **Proportionality to the size of the general-purpose AI model provider –** Measures and KPIs related to the obligations applicable to providers of general-purpose AI models should take due account of the size of the general-purpose AI model provider and allow simplified ways of compliance for SMEs and start-ups with fewer financial resources than those at the frontier of AI development, where appropriate. KPIs related to the obligations applicable to providers of general-purpose AI models with systemic risk shall also reflect differences in size and capacity of providers, where appropriate.

VI. **Support and growth of the AI safety ecosystem –** We recognize that the development, adoption, and governance of general-purpose AI models are a global issue. Many Measures in this draft are intended to enable and support cooperation between the different stakeholders, for example by sharing general-purpose AI safety infrastructure and best practices amongst model providers but also by further enabling the contributions of civil society, academia, third parties and government organisations. This is why we encourage further transparency between stakeholders and increased efforts to share knowledge and cooperate in building a collective and robust evidence base for AI Safety, in line with Article 56(1)(3) and Recital 116 AI Act. We also acknowledge the positive impact that open-source models have had on the development of the AI safety ecosystem.

Table of contents

I. PREAMBLE

Whereas:

a) The Signatories of this Code of Practice (Code) recognise the importance of improving the functioning of the internal market, of creating a level playing field for the regulation of human-centric and trustworthy Artificial Intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter, including democracy, the rule of law and environmental protection, against harmful effects of AI in the Union and supporting innovation as emphasised in Article 1(1) of the AI Act (Act). The Code shall be interpreted in this context.

b) The Code provides guidance on compliance with the obligations set forth in Articles 53 and 55 of the AI Act for providers of general-purpose AI models and general-purpose AI models with systemic risk.

c) Whenever the Code refers to providers of general-purpose AI models it shall encompass providers of general-purpose AI models with systemic risk, too. Whenever the Code refers to providers of general-purpose AI models with systemic risk it shall not encompass providers of other general-purpose AI models.

d) The Signatories recognise that the Code serves as a guiding document for providers of GPAI models and general-purpose AI models with systemic risk in demonstrating compliance with the AI Act, while recognising that adherence to this Code does not constitute conclusive evidence of compliance with the AI Act.

e) The Signatories recognise the importance of reporting their implementation and outcomes of the Code to facilitate the regular monitoring and evaluation of the Code’s adequacy by the AI Office and the Board.

f) The Code shall be subject to regular review by the AI Office. The AI Office may encourage and facilitate updates of the Code to reflect advances in AI technology, societal changes, and emerging systemic risks.

g) The Signatories recognise that the Code serves as a bridge until the adoption of harmonised EU standards for general-purpose AI models. Updates may be needed to facilitate a gradual transition towards future standards.

h) The Signatories recognise that the absence of specific Measures, Sub-Measures, and Key Performance Indicators (KPIs) within this Code does not absolve providers of general-purpose AI models with systemic risk of their responsibility to address and mitigate potential systemic risks as they emerge.

i) The AI Office and Signatories shall work in partnership to foster collaboration between providers of general-purpose AI models, researchers, and regulatory bodies to address emerging challenges and opportunities in the AI landscape.

The Objectives of the Code are as follows:

I. Providers of general-purpose AI models can effectively comply with their obligations. The Code of Practice should clarify to providers how to demonstrate compliance. The Code should also enable the AI Office to assess the compliance of providers who choose to rely on the Code to demonstrate compliance, in accordance with Article 56. This can include allowing sufficient visibility into trends in the development and deployment of general-purpose AI models, particularly of the most advanced models.

II. Providers of general-purpose AI models can effectively ensure a good understanding of general-purpose AI models along the AI value chain, both to enable the integration of such models into downstream products and to fulfil subsequent obligations under the AI Act or other regulations (see Article 53 and Recital 101).

III. Providers of general-purpose AI models can effectively comply with Union law on copyright and related rights (see Article 53 and Recital 106).

IV. Providers of general-purpose AI models with systemic risk can effectively continuously assess and mitigate possible systemic risks at the Union level, including their sources, that may stem from the development, the placing on the market, or the use of general-purpose AI models with systemic risk (see Article 55 and Recital 114).

II. RULES FOR PROVIDERS OF GENERAL-PURPOSE AI MODELS

Whereas:

a) The Signatories recognise the particular role and responsibility of providers of general-purpose AI models along the AI value chain, as the models they provide may form the basis for a range of downstream systems, often provided by downstream providers that need significant understanding of the models and their capabilities, both to enable the integration of such models into their products and to fulfil their obligations under the AI Act.[1]

b) The Signatories recognise that general-purpose AI models and in particular large generative AI models – capable of generating text, images, and other content – present unique innovation opportunities but also challenges to artists, authors, and other creators, and to the way their creative content is created, distributed, used, and consumed. They further recognise that any use of copyright protected content requires the authorisation of the rightsholder(s) concerned unless relevant copyright exceptions and limitations apply.[2]

c) 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.[3]

d) The AI Act and the Code are without prejudice to the rules laid down by Union and national law, and the Code shall be interpreted in particular in accordance with Union copyright law. Directive (EU) 2019/790 introduced exceptions and limitations allowing reproductions and extractions of works or other subject matter, for the purpose of text and data mining, under certain conditions. Under these rules, rightsholders may choose to reserve their rights over their works or other subject matter to prevent text and data mining, unless this is done for the purposes of scientific research. Where rights have been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightsholders if they want to carry out text and data mining over such works.[4]

e) The Signatories recognise that, according to Art. 53(1)(c) AI Act, any provider placing general-purpose AI models on the Union market is obliged to put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790, regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place.[5] Chapter 2 of this Section aims to contribute to the proper application of this obligation[6] by setting out a robust framework to ensure copyright compliance and transparency, while striking a fair balance between the various rights and legitimate

[1] Recital 101.
[2] Recital 105.
[3] Recital 109.
[4] Recital 105.
[5] Recital 106.
[6] Article 56(1).

interests at issue.[7] These Measures also take due account of the interests of small and medium enterprises (SMEs), including startups.

Therefore, the Signatories of this Code commit to the following:

[7] See Articles 17(2), 16 and 13 CFEU and CJEU Judgment of 29 January 2008, Promusicae (C-275/06, ECR 2008 p. I-271) ECLI:EU:C:2008:54, para 68; Judgment of 27 March 2014, UPC Telekabel Wien (C-314/12) ECLI:EU:C:2014:192, para 46; Judgment of 26 April 2022, Poland / Parliament and Council (C-401/19, Publié au Recueil numérique) ECLI:EU:C:2022:297, para 66.

TRANSPARENCY

LEGAL TEXT

Article 53(1)(a): “Providers of general-purpose AI models shall draw up and keep up-to-date the technical documentation of the model, including its training and testing process and the results of its evaluation, which shall contain, at a minimum, the information set out in Annex XI for the purpose of providing it, upon request, to the AI Office and the national competent authorities;”

Article 53(1)(b): “Providers of general-purpose AI models shall draw up, keep up-to-date and make available information and documentation to providers of AI systems who intend to integrate the general-purpose AI model into their AI systems. Without prejudice to the need to observe and protect intellectual property rights and confidential business information or trade secrets in accordance with Union and national law, the information and documentation shall: (i), enable providers of AI systems to have a good understanding of the capabilities and limitations of the general-purpose AI model and to comply with their obligations pursuant to this Regulation; and (ii), contain, at a minimum, the elements set out in Annex XII;”.

Measure 1. Documentation for the AI Office

Signatories commit to draw up and keep up-to-date the technical documentation of the model listed in the table below for the purpose of providing it, upon request, to the AI Office and the national competent authorities. The Signatories are encouraged to consider if the listed information can be disclosed, in whole or in part, to the public to advance public transparency.

Measure 2. Documentation for downstream providers

Signatories commit to draw up, keep up-to-date, and make available information and documentation listed in the table below to providers of AI systems who intend to integrate the general-purpose AI model into their AI systems. The Signatories are encouraged to consider if the listed information can be disclosed, in whole or in part, to the public to advance public transparency.

AI Act reference	**Detailing of information required**	**For the AI Office and national competent authorities**	**For downstream providers**
Annex XI §1 1. and Annex XII 1.	General information: Signatories should detail general information about the provider of the general-purpose AI model and about the model itself to clearly identify and characterise the model, such as the model name, evidence of the provenance and authenticity of the model by means of e.g. a secure hash in the case binaries are distributed or TLS/SSL certificates in the case of a service, legal business name of the developer(s) and the owner(s) of the model in case they are not the same, the model family trade name, the unique name of each model version submitted, etc.	✓	✓

Annex XI §1 1.(a) and Annex XII 1.(a)	Intended tasks and type and nature of AI systems in which it can be integrated: Signatories should provide a description of the intended and restricted or prohibited tasks. This description should also contain the type and nature of AI systems in which the general-purpose AI model can be integrated into, including the high-risk AI applications (as specified in Annex III), if any.	✓	✓
Annex XI §1 1.(b) and Annex XII 1.(b)	Acceptable use policies: Signatories should provide the details of the acceptable use policy (AUP) based on common practices across providers. A valid acceptable use policy should contain, at minimum, the essential elements defined in the Appendix. Signatories should disclose the active URL for the up-to-date acceptable use policy.	✓	✓
Annex XI §1 1.(c) and Annex XII 1.(c)	Date of release and methods of distribution: Signatories should provide the date of release along with an up-to-date list of all the methods of distribution of the general-purpose AI model. Signatories shall disclose the active URL for the up-to-date methods of distribution.	✓	✓
Annex XII 1.(d)	Interaction of the model with external hardware or software: Signatories should provide documentation about how the model interacts with hardware and software, specifying which hardware and which software are not part of the model. Signatories shall disclose versioned dependencies for required software and/or hardware.		✓
Annex XII 1.(e)	Versions of relevant software where applicable: Signatories should provide the details about the versions of the relevant software that is necessary to use the general-purpose AI model. Signatories shall disclose versioned dependencies for required software.		✓
Annex XI §1 1.(d) and Annex XII 1.(f)	Architecture and number of parameters: Signatories should provide a description of the model architecture, the type of model, the context size where appropriate, the total number of model parameters and the number of parameters that are active during inference.	✓	✓
	Signatories should provide greater detail about the model architecture, including the number and types of layers of the model.	✓	
Annex XI §1 1.(e) and Annex XII 1.(g) and 2.(b)	Modality and format of inputs and outputs: Signatories should detail the input and output modalities and associated context limits where applicable.	✓	✓
Annex XI §1 1.(f) and Annex XII 1.(h)	Licence: Signatories should detail the core elements of the licence based on common practices across providers. This includes information about what assets are released (e.g. data, model weights, etc.) and the licence obligations in the	✓	✓

	terms of use, modification and distribution. Signatories shall disclose the active URL for the up-to-date licence.		
Annex XI §1 2.(a) and Annex XII 2.(a)	Technical means for integration into AI systems: Signatories should detail the technical documentation, infrastructure and tools necessary for the general-purpose AI model to be properly integrated in AI systems. Signatories shall disclose versioned dependencies for required software and/or hardware.	✓	✓
Annex XI §1 2.(b)	Design specification and training process: Signatories should detail the core elements of model training (e.g. training stages, the objectives being optimised, the methods of optimization, constraints, etc...), the associated rationale(s) and assumption(s) for the design decisions, and other training details.	✓	
Annex XI §1 2.(c) and Annex XII 2.(c)	Information on data used for training, testing and validation: Signatories should detail the data acquisition methods, specific information for each data acquisition method (e.g. web crawling, data licencing, data annotation, synthetically generated data, user data, etc...), details about the data processing (e.g. if and how harmful or private data are filtered), and specific information about the data used to train/test/validate the model, such as the fraction of the data that comes from different data sources, and the main characteristics of the training, testing and validation data.	✓	✓
	Signatories should further detail the size (number of data points) of the training, test, and validation data for each data modality (e.g. text, images, videos) and the methods used to detect unsuitability of data sources and any biases in the data.	✓	
Annex XI §1 2.(d)	Computational resources: Signatories should detail the computational resources (e.g. the number and type of hardware units needed to train and do inference with the general-purpose AI model, the duration of the training process, the number of FLOPs) used to train and do inference with the model, in consistency with any delegated act adopted in accordance with Article 97 of the AI Act to detail measurement and calculation methodologies with a view to allowing for comparable and verifiable documentation.	✓	
Annex XI §1 2.(e)	Energy consumption: Signatories should detail which information and methodology they use to assess energy consumption (e.g. hardware provider, location and energy sources associated with the hardware, energy consumed, estimated emissions generated), in consistency with any delegated act adopted in accordance with Article 97 of the AI Act to detail measurement and calculation methodologies with a view to allowing for comparable and verifiable documentation.	✓	

Article 53 (1) (a)	Testing process and results thereof: Signatories should detail the testing process they conduct of the general-purpose AI model, including if no testing is conducted. These details should include a description of the tests performed and the results of these tests to ensure proper interpretation.	✓	

OPEN QUESTION

For the items listed in the table above, how should the Code provide greater detail?

Appendix: Essential elements of an Acceptable Use Policy

An Acceptable Use Policy (AUP) is defined as a set of rules that outline how a service or technology can be used. It is a document that provides guidelines to users on what is and isn’t acceptable behaviour. The AUP should be consistent with the Signatories' materials that describe the uses and capabilities of their general-purpose AI model. The Signatories should commit to sharing with the downstream providers all the necessary information related to their general-purpose AI model to enable downstream providers to comply with existing regulations applicable to the task or use case their AI system is intended to be used for.

The AUP should, at the very least, include:

- A purpose statement explaining why the AUP exists;
- The scope defining who the policy applies to and what resources it covers;
- Primary intended uses and users;
- Acceptable uses, listing activities and tasks that are allowed, including high-risk AI applications (as specified in Annex III), if any, the model is intended to be integrated into;
- Unacceptable uses, detailing forbidden actions;
- Security measures containing a description of the security protocols that the users of the general-purpose AI systems must follow;
- Monitoring and privacy, explaining why and how the general-purpose AI provider monitors the use of their model and impact on user’s privacy;
- Warning processes and criteria for suspension or withdrawal of user privileges for not adhering to the AUP;
- Criteria for terminating user accounts and reference to applicable law and regulations for enforcement;
- Acknowledgement, requiring downstream providers to acknowledge that they have read, understood and agreed to comply with the AUP.

RULES RELATED TO COPYRIGHT

LEGAL TEXT

Article 53(1)(c): “Providers of general-purpose AI models shall put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790;”.

Measure 3. Put in place a copyright policy

Signatories commit to put in place a policy to comply with Union law on copyright and related rights.

In order to satisfy Measure 3:

Sub-Measure 3.1. Draw up and implement a copyright policy

Signatories will draw up and implement an internal policy to comply with Union law on copyright and related rights in line with this Chapter of the Code. This policy shall cover the entire lifecycle[8] of any general-purpose AI model subject to this Code. In the case of a modification or fine-tuning of a general-purpose AI model, the obligations for providers of general-purpose AI models only concern that modification or fine-tuning, including new training data sources, in accordance with guidance provided by the AI Office.[9] Signatories will assign responsibilities within their organisation for the implementation and overseeing of this policy.

Sub-Measure 3.2. Upstream copyright compliance

Signatories will undertake a reasonable copyright due diligence before entering into a contract with a third party about the use of data sets for the development of a general-purpose AI model. In particular, Signatories are encouraged to request information from the third party about how the third party identified and complied with, including through state-of-the-art technologies, rights reservations expressed pursuant to Article 4(3) of Directive (EU) 2019/790.

Sub-Measure 3.3. Downstream copyright compliance

Signatories will implement reasonable downstream copyright measures to mitigate the risk that a downstream system or application, into which a general-purpose AI model is integrated, generates copyright infringing output.[10] The downstream copyright policy should take into account whether a signatory vertically integrates an own general-purpose AI model into its own AI system or whether a general-purpose AI model is provided to another entity based on contractual relations.[11] In particular, signatories are encouraged to avoid an overfitting of their general-purpose AI model and to make the conclusion or validity of a contractual provision of a general-purpose AI model to another entity dependent

[8] Recital 65.
[9] Recital 109.
[10] Article 3(1) and Article 3(63).
[11] Recital 97 and Article 3(68).

upon a promise of that entity to take appropriate measures to avoid the repeated generation of output which is identical or recognisably similar to protected works. This Sub-Measure does not apply to SMEs.

Measure 4. Compliance with the limits of the TDM exception

When Signatories engage in text and data mining according to Article 2(2) of Directive (EU) 2019/790 for the development of their general-purpose AI models, they commit to ensure that they have lawful access to copyright-protected content and to identify and comply with rights reservations expressed pursuant to Article 4(3) of Directive (EU) 2019/790.

In order to satisfy Measure 4:

Sub-Measure 4.1. Respect Robots.txt

Signatories will only employ crawlers that read and follow instructions expressed in accordance with the Robot Exclusion Protocol (robots.txt).

Sub-Measure 4.2. No effect on findability

Signatories that also provide an online search engine as defined in Article 3(j) Regulation (EU) 2022/2065 or control such a provider will take appropriate measures to ensure that a crawler exclusion expressed pursuant to the Robot Exclusion Protocol does not negatively affect the findability of the content in their search engine.

Sub-Measure 4.3. Best efforts regarding other appropriate means

Signatories will make best efforts in accordance with widely used industry standards to identify and comply with other appropriate machine-readable means to express a rights reservation at source and/or work level pursuant to Article 4(3) of Directive (EU) 2019/790 in the case of content made publicly available online. In particular, Signatories are encouraged to implement widely adopted tools that enable expressions of rights reservations at aggregate level.

Sub-Measure 4.4. Commitment to collaborative development of rights reservations' standards

Upon an invitation by the Commission, Signatories will engage in bona fide discussions with entities that are sufficiently representative of affected rightsholders, and with other relevant stakeholders such as standardisation organisations, to develop interoperable machine-readable standards to express a rights reservation pursuant to Article 4(3) of Directive (EU) 2019/790 and to identify and comply with such a rights reservation. The Commission will convene and chair meetings, and after consultations with relevant stakeholders and general-purpose AI providers, as appropriate, may issue information about state-of-the-art solutions that providers are expected to honour. This Sub-Measure does not apply to SMEs. SMEs may, however, voluntarily participate in these discussions.

Sub-Measure 4.5. No crawling of piracy websites

Signatories will take reasonable measures to exclude pirated sources from their crawling activities, such as by excluding websites listed in the Commission Counterfeit and Piracy Watch List. Signatories are also encouraged to comply with analogous exclusion lists published by relevant public authorities in the jurisdictions where they are established.

Measure 5. Transparency

Signatories commit to adequate transparency about the measures they adopt to comply with Union law on copyright and related rights.

In order to satisfy Measure 5:

Sub-Measure 5.1. Public information about rights reservation compliance

Signatories will make public, in a language broadly understood by the largest possible number of Union citizens, adequate information about the measures they adopt to identify and comply with rights reservations expressed pursuant to Article 4(3) of Directive (EU) 2019/790. That information shall be easily accessible on each Signatory's website and shall be kept up to date.

Sub-Measure 5.2 Crawler name and robots.txt features

The information according to the preceding Sub-Measure includes, at a minimum, the name of all crawlers that the signatory uses for the development of a general-purpose AI model subject to this Code and their relevant robots.txt features, including at the time of crawling.

Sub-Measure 5.3. Single point of contact and complaint handling

Signatories are encouraged to designate a single point of contact to enable rightsholders to communicate directly and rapidly with them, by electronic means. In particular, they are encouraged to enable rightsholders and their representatives, including collective management bodies, to lodge complaints concerning the use of their works or other protected subject matter for the development of a general-purpose AI model, and to implement appropriate complaint handling procedures.

Sub-Measure 5.4 Documentation of data sources and authorisations

In order to allow the AI Office to monitor[12] whether Signatories have fulfilled the obligation to put in place a policy to comply with Union law on copyright and related rights,[13] Signatories will draw up, keep up-to-date and provide the AI Office upon its request with information about data sources used for training, testing and validation and about authorisations to access and use protected content for the development of a general-purpose AI.

[12] Article 89(1).
[13] Recital 108 and Article 53(1).

III. TAXONOMY OF SYSTEMIC RISKS

Whereas:

a) The Signatories recognise that the taxonomy of systemic risks includes types, nature, and sources of systemic risks.

b) The Signatories recognise that the taxonomy has been developed and, when in doubt, should be interpreted in light of the severity and probability of each risk as defined in Article 3(2) of the AI Act and of the definition of systemic risk as defined in Article 3(65) of the AI Act.

c) 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.

d) The Signatories recognise that Section III and Section IV generally refer to general-purpose AI models and not AI systems but that some risks are often best identified, assessed, evaluated, and mitigated by taking into account how the general-purpose AI model could be deployed in AI systems. In cases where general-purpose AI model providers also develop and operate AI systems based on general-purpose AI models with systemic risk, they will do risk assessment and mitigation (as described in the Safety and Security Framework) by taking into account these systems.

Therefore, the Signatories of this Code commit to the following:

Measure 6. Taxonomy

Signatories commit to draw from the elements of this taxonomy of systemic risks as a basis for their systemic risk assessment and mitigation.

6.1. Types of systemic risks

Signatories will treat the following as systemic risks:

- **Cyber offence**: Risks related to offensive cyber capabilities such as vulnerability discovery or exploitation.
- **Chemical, biological, radiological, and nuclear risks**: Dual-use science risks enabling chemical, biological, radiological, and nuclear weapons attacks via, among other things, weapons development, design, acquisition, and use.
- **Loss of Control:** Issues related to the inability to control powerful autonomous general-purpose AI models.
- **Automated use of models for AI Research and Development:** This could greatly increase the pace of AI development, potentially leading to unpredictable developments of general-purpose AI models with systemic risk**.**
- **Persuasion and manipulation:** The facilitation of large-scale persuasion and manipulation, as well as large-scale disinformation or misinformation with risks to democratic values and human rights, such as election interference, loss of trust in the media, and homogenisation or oversimplification of knowledge.
- **Large-scale discrimination:** Large-scale illegal discrimination of individuals, communities, or societies.

Signatories may identify further systemic risks beyond those listed above, considering, for example, major accidents, large-scale privacy infringements and surveillance, as well as other ways in which general-purpose AI models may cause large-scale negative effects on public health, safety, democratic processes, public and economic security, critical infrastructure, fundamental rights, environmental resources, economic stability, human agency, or society as a whole.

OPEN QUESTIONS

- What are relevant considerations or criteria to take into account when defining whether a risk is a systemic risk?
- Based on these considerations or criteria, which risks should be prioritised for addition to the main taxonomy of systemic risks?
- How should the taxonomy of systemic risks address AI-generated child sexual abuse material and non-consensual intimate imagery?

6.2. Nature of systemic risks

The nature of systemic risks refers to key attributes of risks that influence how these may be assessed and mitigated. Signatories consider the below particularly relevant dimensions of the nature of systemic risks and examples for each dimension that are neither exhaustive nor mutually exclusive:

- **Origin**: Model capabilities, model distribution
- **Actor(s) driving the risk:** State, group, individual, autonomous AI agent, none (e.g., no clear actor can be identified)
- **Intent:** Intentional, unintentional (including misalignment)
- **Novelty:** Precedented, unprecedented
- **Probability-severity ratio:** Low-impact high-probability, high-impact low-probability, high expected impact
- **Velocity at which the risk materialises:** Gradual, sudden, continuously changing
- **Visibility of the risk while it materialises:** Overt (open), covert (hidden)
- **Course of events:** Linear, recursive (feedback loops), compound, cascading (chain reactions)

6.3. Sources of systemic risks

Sources of risks, also referred to as “factors of risks” or “drivers of risks”, are elements (e.g. events, components, actors and their intentions or activities) that alone or in combination give rise to risks (e.g. model theft or widespread cyber vulnerabilities). Signatories consider the below particularly relevant sources of systemic risks:

6.3.1. Dangerous model capabilities

These are model capabilities that may cause systemic risk. Signatories recognise that many of these capabilities are also important for beneficial uses. These include:

- Cyber-offensive capabilities, Chemical, Biological, Radiological and Nuclear (CBRN) capabilities, and weapon acquisition or proliferation capabilities

- Autonomy, scalability, adaptability to learn new tasks
- Self-replication, self-improvement, and ability to train other models
- Persuasion, manipulation, and deception
- Long-horizon planning, forecasting, and strategising
- Situational awareness

6.3.2 Dangerous model propensities

These are model characteristics beyond capabilities that may cause systemic risk. They include:

- Misalignment with human intent and/or values
- Tendency to deceive
- Bias
- Confabulation
- Lack of reliability and security
- “Goal-pursuing”, resistance to goal modification, and “power-seeking”
- “Colluding” with other AI models/systems to do so

6.3.3 Model affordances and socio-technical context

These are factors beyond model capabilities and propensities that may influence the systemic risks posed by the model. They encompass specific inputs, configurations, and contextual elements of a general-purpose AI model with systemic risk. These include:

- Potential to remove guardrails
- Access to tools (including other models)
- Modalities (including novel and combined modalities)
- Release and distribution strategies
- Human oversight
- Model exfiltration (e.g. model leakage/theft)
- Number of business users and number of end-users
- Offence-defence balance, including the number, capacity, and willingness of bad actors to misuse the model
- Societal vulnerability or adaptation
- Lack of explainability or transparency
- Technology readiness (i.e. how mature a technology is within a given application context)

Feedback loops in the use of data, model, and inferences

IV. RULES FOR PROVIDERS OF GENERAL-PURPOSE AI MODELS WITH SYSTEMIC RISK

EXPLANATORY BOX

The Measures, Sub-Measures, and KPIs should be proportionate. In particular, they will require a tailoring to the size and capacity of a specific provider, particularly SMEs and start-ups with less financial resources than those at the frontier of AI development, and to different distribution strategies (e.g. open-sourcing), where appropriate, reflecting the principle of proportionality and taking into account both benefits and risks.

The current draft is written under the assumption that there will be a small number of both general-purpose models with systemic risks and providers thereof. Future drafts may need to be changed significantly should these numbers grow, for instance by introducing a more detailed tiered system of Measures aiming to focus primarily on those models that pose the largest systemic risks.

The "whereas" part immediately below is a preamble for Section IV. Here, high-level principles guide the interpretation of the Measures, Sub-Measures, and KPIs.

Finally, this is only our first draft. We are looking forward to your feedback. We have highlighted relevant open questions, but welcome input on other parts of the draft as well. We also welcome suggestions on how the Measures can be made more proportionate, as well as more appropriate, for different business models and deployment strategies.

Chairs and Vice-Chairs of Working Groups 2, 3, and 4.

LEGAL TEXT

Article 55(1): "In addition to the obligations listed in Articles 53 and 54, providers of general-purpose AI models with systemic risk shall:

(a) perform model evaluation in accordance with standardised protocols and tools reflecting the state of the art, including conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risks;

(b) assess and mitigate possible systemic risks at Union level, including their sources, that may stem from the development, the placing on the market, or the use of general-purpose AI models with systemic risk;

(c) keep track of, document, and report, without undue delay, to the AI Office and, as appropriate, to national competent authorities, relevant information about serious incidents and possible corrective measures to address them;

(d) ensure an adequate level of cybersecurity protection for the general-purpose AI model with systemic risk and the physical infrastructure of the model."

Whereas:

a) Signatories recognise that providers of general-purpose AI models with systemic risk should continuously assess and mitigate systemic risks, taking appropriate measures along the entire model's lifecycle, cooperating with relevant actors along the AI value chain, and ensuring their risk management is future-proof by regularly updating their practices in light of improving and emerging capabilities.[14]

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, (ii) has uncertain capabilities and impacts, or (iii) where the provider lacks relevant expertise. Conversely, there is less need for more comprehensive measures where there is good reason to believe that a new general-purpose AI model will exhibit the same high-impact capabilities as exhibited by general-purpose AI models with systemic risk that have already been safely deployed, without significant systemic risks materialising and where the implementation of appropriate mitigations has been sufficient. 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.

c) Signatories recognise that there are a wide range of organisations that have significant expertise and are well placed to assist with the assessment and mitigation of systemic risks.

d) Signatories recognise that many risk assessment methods come with significant workload and costs. They encourage each other to 'share the load', for example by sharing evaluations, best practices or infrastructure, or – where appropriate – by working with qualified third-party providers, potentially facilitated by industry organisations.

e) Signatories intend to interpret the Measures, Sub-Measures, and KPIs, when in doubt, in light of the effective assessment and mitigation of systemic risks.

Therefore, the Signatories of this Code commit to the following:

Measure 7. Safety and Security Framework

The Signatories commit to adopting, implementing, and making available a Safety and Security Framework (SSF), which shall detail the risk management policies they adhere to in order to proactively assess and proportionately mitigate systemic risks from their general-purpose AI models with systemic risks (see Article 55(1)). The comprehensiveness of an SSF as well as the commitments in it should be proportional to the severity of expected systemic risks from the development of such models. The initial required draft components of an SSF are outlined in the remainder of this section.

[14] Recital 114 AI Act ("providers of general-purpose AI models with systemic risks should continuously assess and mitigate systemic risks, including for example by putting in place risk-management policies, such as accountability and governance processes, implementing post-market monitoring, taking appropriate measures along the entire model's lifecycle and cooperating with relevant actors along the AI value chain").

RISK ASSESSMENT FOR PROVIDERS OF GENERAL-PURPOSE AI MODELS WITH SYSTEMIC RISK

Measure 8. Risk identification

As part of their SSF, Signatories commit to continuously and thoroughly identifying systemic risks that may stem from the general-purpose AI model with systemic risk.

In order to satisfy Measure 8.:

Sub-Measure 8.1. Determining risks

Signatories will determine and specify the systemic risks that are particularly relevant to the proposed development, placing on the market, or use of the general-purpose AI model with systemic risk. For this purpose, they will use the systemic risks listed in the taxonomy (Section III) and may consider additional risks as well as reference other elements of the taxonomy.

Measure 9. Risk analysis

As part of the SSF, Signatories commit to carrying out a continuous and thorough analysis of the pathways to systemic risks identified.

In order to satisfy Measure 9:

Sub-Measure 9.1. Methodologies

Signatories will use robust risk analysis methodologies to identify the pathways by which the development and deployment of their general-purpose AI model with systemic risk could produce the systemic risks identified, as well as the probability of such risks materialising through those pathways.

Sub-Measure 9.2. Mapping to systemic risk indicators

For their general-purpose AI models with systemic risk, Signatories will identify and map potentially dangerous model capabilities, propensities, and other sources of risk that may enable the pathways to systemic risks identified**,** and provide systemic risk indicators for each of these elements.

Sub-Measure 9.3. Tiers of severity

For their general-purpose AI models with systemic risk, Signatories will categorise the identified dangerous model capabilities, dangerous model propensities, and other sources of risk into tiers of severity, including at minimum a tier of severity at which the level of risk would be considered intolerable absent appropriate safeguards.

OPEN QUESTIONS

- What will those tiers of severity be? Is there already a nascent standard or a consensus forming?
- Is “severity” the best way to articulate levels of “gravity” or could it create confusion with the definition of risk as the combination of probability and severity?

Sub-Measure 9.4. Forecasting risks

Signatories will include in their SSF best effort estimates of timelines for when they expect to develop a model that triggers the systemic risk indicators mentioned in Sub-Measure 9.2.

Measure 10. Evidence Collection

As part of the SSF, Signatories commit to a continuous process of Evidence Collection on the specific systemic risks presented by their general-purpose AI models with systemic risk. 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.

In order to satisfy Measure 10.:

Sub-Measure 10.1. Model-agnostic evidence

Where applicable to their general-purpose AI models with systemic risk, Signatories will collect model-agnostic evidence of the systemic risks presented by their model, using a wide range of methods that may include literature reviews, competitor and open-source project analysis, forecasting of general trends (like algorithmic efficiency, compute use, energy use, etc.), and participatory methods involving civil society, academia, and other relevant stakeholders. They may also work on scaling laws that predict capability gains from scaling models. Like all evidence collection in this section, this may be done in collaboration with – or outsourced to – qualified third parties.

Sub-Measure 10.2. Best-in-class evaluations

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. This shall happen at the most appropriate times during the lifecycle of an AI model with systemic risk, using a range of suitable methodologies (for example Q&A sets, benchmarks, red-teaming and other methods of adversarial testing, human uplift studies, model organisms, simulations, and proxy evaluations for classified materials), and shall be done by evaluators (internal or external) qualified for the relevant risk. The depth of these evaluations shall be proportional to the risk being evaluated and to the uncertainty about how much risk such a model adds (existing knowledge about the behaviour of very similar models may for example reduce the depth of evaluation needed).

OPEN QUESTIONS

What factors might determine whether a certain evaluation method is an adequate fit for a specific model and risk, and whether an evaluation was thorough enough?

Sub-Measure 10.3. Scientific rigour and other quality factors

Signatories will ensure the execution of evaluations with high scientific rigour. Additional rigour shall be achieved through the validation of key results by qualified third parties, especially for high tiers of severity of systemic risks (see Measure 17)**.** Signatories will provide evaluators, internal or external, with the support needed to work to a rigorous scientific standard, including enough time, model access, and compute budget to properly evaluate a general-purpose AI model with systemic risk, while protecting intellectual property rights and confidential business information where appropriate.

OPEN QUESTIONS

How should high scientific rigor be operationalised? What is the gold standard and when should Signatories deviate from it (for example when conducting early, exploratory research)?

Sub-Measure 10.4. Capability elicitation

Signatories will ensure that evaluations are being run with a best-in-class level of capability elicitation (e.g. fine-tuning, prompt engineering, scaffolding, compute and engineering budgets) to fully elicit the capabilities of a model and minimise the risk of under-estimating capabilities.

Sub-Measure 10.5. Models as part of systems

Signatories will ensure that evaluations can assess the capabilities and limitations of a general-purpose AI model with systemic risk, both in an AI system representative of future AI systems in which the model is intended to and reasonably foreseeably will be used, but also in an AI system where the model's maximum potential to pose systemic risks is revealed.

OPEN QUESTION

How could this Sub-Measure be facilitated for Signatories who provide general-purpose AI models with systemic risk as open-source models or to Business-to-Business customers?

Sub-Measure 10.6. Diverse evaluations & generalisation

Signatories will ensure that evaluations match the planned usage context of a model with all its variety, where applicable, to show generalisation. For example, language-based evaluations of multilingual models may focus not only on English, but on multilingual evaluations that take into account European diversity.

Sub-Measure 10.7. Exploratory work

Signatories will ensure that significant amounts of exploratory work are done on their general-purpose models with systemic risk, such as open-ended red teaming by qualified third parties (including representatives of civil society and academia). This means that they will not restrict themselves only to evidence collection for risks or capabilities they have already identified, but also strive to identify new risks and emerging capabilities through these methods.

Sub-Measure 10.8. Sharing tools & best practices

Signatories will strive to make best-in-class safety evaluations, tooling, and accompanying best practices widely accessible to relevant actors in the AI ecosystem. In specifically identified cases, Signatories may limit the sharing of information to protect commercially sensitive information, public security, proliferation risks, and the validity of future evaluations.

OPEN QUESTIONS

- What channels, organisations and methods exist that would facilitate the sharing of evaluations, tools, and best practices, while not putting undue additional pressure on the research teams currently working at the cutting edge of AI Safety?
- Is this measure especially beneficial to startups and SMEs who might not have as much capacity to develop these tools and practices from scratch, but might be able to use them?

Sub-Measure 10.9. Sharing results

When Signatories share evaluation results with the AI Office or the public, they will do so in a transparent and easily comparable format. They shall transparently report uncertainty of any empirical results and limitations of the methods used.

Measure 11. Risk assessment lifecycle

Signatories commit to continuously assess risks and collect evidence during the full lifecycle of the development and deployment of general-purpose AI models with systemic risk, at least at the stages outlined in the Sub-Measures to this Measure and before and after implementing mitigations (including assessing the effectiveness of mitigations outlined in Measure 12).

In order to satisfy Measure 11.:

Sub-Measure 11.1. Before training

Before starting a training run for a general-purpose AI model with systemic risk, Signatories will make updates to the SSF as necessary and ensure evaluators (internal and external) are ready for Evidence Collection, in line with Signatories' SSF commitments.

Sub-Measure 11.2. During training

Signatories will collect evidence at regular milestones (as an example, this could be every four-fold increase in effective compute), updating an in-progress Safety and Security Report (SSR, see Measure 13) as commensurate with the risks. Training here is not restricted to mean only "pre-training on a large corpus of data" but shall also include, for example, supervised fine tuning, reinforcement learning phases, or similar methods of refining a model.

Sub-Measure 11.3. During deployment

During the deployment of any general-purpose AI model with systemic risk, Signatories will update the model's SSR by revisiting their risk assessment, especially by re-running relevant evaluations (and/or newer and improved evaluations) at least every six months, or whenever they perceive a major change of (internal or external) circumstances or have reason to doubt previous risk assessment results for any other reason, also taking into account any evidence gained from monitoring of the model in deployment.

Sub-Measure 11.4. Post-deployment monitoring

Signatories will conduct post-deployment monitoring for systemic risks. They will establish mechanisms to continuously gather and include relevant post-deployment information in risk assessment. These mechanisms may vary across different model integrations and usage (for example monitoring of models for harmful outputs and actions or investigating systemic impacts). Signatories will adapt their post-deployment monitoring to the distribution strategy and the type of customers and industries using the model (e.g. for open weight models they may consider evaluating adherence to licenses, monitor evidence of model usage in the real world, or study scientific analyses of the model). Where model providers themselves deploy AI systems, they will monitor these models as part of these systems.

OPEN QUESTIONS

What methods exist (or could exist) that would enable providers of open-weights general-purpose AI models with systemic risk to monitor models they have released, without major side effects for the downstream users of these models?

TECHNICAL RISK MITIGATION FOR PROVIDERS OF GENERAL-PURPOSE AI MODELS WITH SYSTEMIC RISK

Measure 12. Mitigations

Signatories commit to detail in their SSF a mapping from each systemic risk indicator or tier of severity to proportionally necessary safety and security mitigations, building on AI Office guidance where available. The mapping should at a minimum be designed to keep systemic risks below an intolerable level, and should also describe how risk will be minimised beyond that.

In order to satisfy Measure 12.:

Sub-Measure 12.1. Safety mitigations

Signatories will detail in their SSF the safety mitigations they will implement to mitigate systemic risk resulting from the use of their general-purpose AI model with systemic risk. These safety mitigations should be proportional to systemic risk indicators or tiers of severity, and could entail (a) behavioural modifications to a model, (b) safeguards placed around the model for deployment in a system, and (c) countermeasures or other safety tools made available to other actors to reduce systemic risk.

Sub-Measure 12.2. Security mitigations

Signatories will detail in their SSF the security mitigations they will implement to mitigate systemic risk from the possession of (a) the unreleased weights of a general-purpose AI model with systemic risk, and (b) related unreleased assets and information necessary to train or use such unreleased models. For unreleased models these security mitigations should apply during the development stage, before sufficient risk assessment has been undertaken to justify a deployment decision. For released closed models these security mitigations should also apply during and after the deployment of the model, but such mitigations need not apply for models with openly released weights or related assets. These security mitigations should furthermore be proportional to systemic risk indicators or tiers of severity, and could entail (a) protection of weights and assets at-rest, in-motion, and in-use, including at the hardware-level as appropriate (b) access control, monitoring, and hardened interfaces to weights and assets, (c) assurance through ongoing security red-teaming and accredited security reviews, and (d) screening for insider threats.

OPEN QUESTIONS

- What standards for cybersecurity and information security should be applied to general-purpose AI models with systemic risks, depending on the systemic risk indicators and tiers of severity?
- In what ways should cybersecurity standards for general-purpose AI models with systemic risk be different from existing cybersecurity standards in other domains?

Sub-Measure 12.3. Limitations

Signatories will detail in their SSF the limitations of existing safety and security mitigations, and state where appropriate mitigations to manage systemic risk do not yet exist for a given systemic risk indicator or tier of severity.

Sub-Measure 12.4. Process for assessing adequacy of mapping

Signatories will detail in their SSF their process for assessing the continued adequacy of their mapping from systemic risk indictors or tiers of severity to safety and security mitigations in Sub-Measures 12.1.—12.2. This should be done to keep pace with changes in internal and external factors relevant to a model's impact, such as advances in capability elicitation and the cybersecurity landscape, going beyond the overall process outlined in Measure 17. for assessing the adequacy of the SSF as a whole

Measure 13. Safety and Security Reports

As part of risk mitigation and assessment, to ensure comparable and verifiable documentation of the Measures 8.—12., Signatories commit to create a Safety and Security Report (SSR) for any general-purpose AI model with systemic risk that they develop. This report shall (a) be done at appropriate decision points in the model development and deployment lifecycle, (b) detail risk and mitigation assessments for the model, and (c) form the basis of any development and deployment decisions for the model.

In order to satisfy Measure 13.:

Sub-Measure 13.1. Proportionality

Signatories will ensure an SSR's (a) comprehensiveness and level of detail, (b) timing in the development and deployment lifecycle, and (c) level of external input and scrutiny, are all proportional to the systemic risk indicators or tiers of severity relevant to the model under assessment.

Sub-Measure 13.2. Results of risk assessment

Signatories will detail in an SSR the results of risk assessment undertaken for the model, both before and after mitigations have been implemented, in line with Measures 8.–11.

Sub-Measure 13.3. Results of safety mitigations assessment

Signatories will detail in an SSR the results of assessments of the effectiveness of implemented safety mitigations in line with Sub-Measure 12.1.

Sub-Measure 13.4. Results of security mitigations assessment

Signatories will detail in an SSR the results of assessments of the effectiveness of implemented security mitigations in line with Sub-Measure 12.2.

Sub-Measure 13.5. Cost-benefit analysis

Signatories will detail in an SSR any cost-benefit analysis used to justify a deployment proceeding according to Sub-Measure 14.2.

Sub-Measure 13.6. Sufficient detail on methodology

Signatories will ensure an SSR has sufficient scientific detail to allow for the independent assessment of the methods used to generate the results, evidence, and analysis in Sub-Measures 13.2.–13.5 (see also Sub-Measure 10.3).

Sub-Measure 13.7. Review

Signatories will detail in an SSR the results from an internal (or external at higher levels of severity) review of the results provided in Sub-Measures 13.2.–13.6.

Sub-Measure 13.8. Equivalency

Signatories will ensure that any SSR shared with the AI Office is the same as what is used internally for a development or deployment decision.

Measure 14. Development and deployment decisions

To mitigate risks from insufficient safety and security mitigations, Signatories commit to establish a process to decide whether to proceed or not with the development and deployment of a general-purpose AI model with systemic risk. This process shall be described in the Signatories' SSF, and shall be based on the results and analysis presented in an SSR.

In order to satisfy Measure 14.:

Sub-Measure 14.1. Conditions for not proceeding

Signatories will detail in their SSF the conditions under which further development and deployment of a general-purpose AI model with systemic risk will not proceed, or an existing general-purpose AI model with systemic risk will be removed from deployment or deleted, based on the SSR for the model after safety and security mitigations have been implemented.

Sub-Measure 14.2. Conditions for proceeding

Signatories will detail in their SSF the conditions under which development or deployment can continue when otherwise not proceeding per Sub-Measure 14.1., such as through the implementation of better safety and security mitigations or through the presentation of a cost-benefit analysis, with a rigour and assessment process for these appropriate to the systemic risk indicators or tiers of severity.

Sub-Measure 14.3. External input and decision-making

Signatories will detail in their SSF when development and deployment decisions will have input or require authorisation from external actors, including relevant government actors such as the AI Office.

GOVERNANCE RISK MITIGATION FOR PROVIDERS OF GENERAL-PURPOSE AI MODELS WITH SYSTEMIC RISK

Measure 15. Systemic risk ownership

Signatories commit to ensure adequate ownership regarding systemic risk at all organizational levels, including at the executive and board levels, so as to assess and proportionally mitigate systemic risks (see Art. 55(1) and Recital 114).

In order to satisfy Measure 15.:

Sub-Measure 15.1. Executive level

Signatories will allocate responsibility and resources at the executive level, for addressing systemic risks produced by their general-purpose AI models with systemic risk.

Sub-Measure 15.2. Board level

Signatories will allocate responsibility and resources for oversight of the systemic risks produced by their general-purpose AI models with systemic risk at the board level (or equivalent), such as by establishing a risk committee.

OPEN QUESTIONS

- Should the above Sub-Measures be made relative to provider size or other relevant characteristics? If so, how?
- Should there be more, or other, examples of what might qualify as adherence to Measure 15?

Measure 16. Adherence and adequacy assessment

Signatories commit to assess the adherence to and adequacy of their SSF (see Article 55(1) and Recital 114).

In order to satisfy Measure 16.:

Sub-Measure 16.1. Periodic SSF assessment

Signatories will conduct and document an annual assessment of both the adequacy of and adherence to their SSF, considering their planned activities, presenting it to the board or equivalent.

OPEN QUESTIONS

- Are there specific questions such an assessment should answer?
- How should adequacy be defined in this context?

Measure 17. Independent expert systemic risk and mitigation assessments

Signatories commit to enable meaningful independent expert risk and mitigation assessment of general-purpose AI models with systemic risk throughout their lifecycle, as appropriate, especially for high severity tiers. Such independent expert risk and mitigation assessment may involve independent testing of model capabilities, reviews of evidence collected, systemic risks, and the adequacy of mitigations. It may also involve independent expert review of the SSF and SSR (see Article 55(1) and Recital 114).

OPEN QUESTIONS

- Under what circumstances is independent expert systemic risk assessment of a general-purpose AI model with systemic risk appropriate before deployment? What about assessment of mitigations? Under what conditions does it seem counterproductive or unnecessary?
- Are there circumstances under which it is appropriate or advisable to involve independent experts in risk assessments iteratively, throughout the lifecycle, starting before or during training?
- How can independent systemic risk assessments be adapted to the magnitude and nature of the relevant systemic risk, e.g. with regards to information security, depth of access to general-purpose AI models with systemic risk components and documentation, scope of testing, time to test, expertise, and transparency?
- How should the measures be made relative to severity levels?

In order to satisfy Measure 17.:

Sub-Measure 17.1. Before deployment

Signatories will ensure sufficient independent expert testing before deployment of general-purpose AI models with systemic risk, such as by the AI Office and appropriate third-party evaluators, in accordance with AI Office guidance where available, to more accurately assess risks and mitigations, and to provide assurance to external actors. This may also include a review of appropriate elements of the evidence collected by the Signatory.

OPEN QUESTION

What constitutes an appropriate third-party evaluator? How can the Code be drafted so as to take into account the current immaturity of the industry? Is there some way providers, especially SMEs, can be supported by the AI Office in ensuring independent expert assessment of risks and mitigations?

Sub-Measure 17.2. After deployment

Signatories will enable meaningful independent testing of general-purpose AI models with systemic risk after deployment, as appropriate, to for example assess risk throughout the model lifecycle and identify suitable post-deployment modifications. This could include allowing independent researchers as well as other relevant parties including the AI Office to meaningfully study the risks, limitations, and properties of models, by for example providing them with sufficient access, resources, and assurances of non-retaliation against legitimate research activity.

OPEN QUESTION

When are different means of facilitating independent testing – such as research safe harbors and vulnerability reporting – appropriate?

Measure 18. Serious incident reporting

LEGAL TEXT

Article 55(1)(c): “In addition to the obligations listed in Articles 53 and 54, providers of general-purpose AI models with systemic risk shall keep track of, document, and report, without undue delay, to the AI Office and, as appropriate, to national competent authorities, relevant information about serious incidents and possible corrective measures to address them;”.

Signatories commit to identify and keep track of serious incidents, as far as they originate from their general-purpose AI models with systemic risk, document and report, without undue delay, any relevant information and possible corrective measures to the AI Office and, as appropriate, to national competent authorities.

OPEN QUESTIONS

- What does a serious incident entail? Should the Code use the definition the AI Act uses for AI systems in Article 3(49) or is another definition more appropriate for general-purpose AI models with systemic risk?
- Under what conditions should a general-purpose AI model with systemic risk be judged to have indirectly led to a serious incident occurring?
- Are there suitable technical standards or best practices that can enable automated or streamlined reporting of serious incidents to the AI Office?

In order to satisfy Measure 18.:

Sub-Measure 18.1. Serious incident reporting processes

Signatories will set up processes (including by designating staff members) to identify, document, and report serious incidents and near-misses to the AI Office, as far as they originate from their general-purpose AI model with systemic risk.

Sub-Measure 18.2. Response readiness

Signatories will set up processes for responding to serious incidents, including pre-defining corrective measures that may be taken in response to serious incidents, along with an explanation of when they may be taken.

OPEN QUESTIONS

- What possible corrective measures could be taken in response to serious incidents? Should the Code specify when they may be appropriate?
- What serious incident response processes are appropriate for open weight or open-source providers?

Measure 19. Whistleblowing protections

LEGAL TEXT

Article 87: "Directive (EU) 2019/1937 shall apply to the reporting of infringements of this Regulation and the protection of persons reporting such infringements."

Signatories commit to implement whistleblowing channels and afford appropriate whistleblowing protections to covered persons and activities.

In order to satisfy Measure 19.:

Sub-Measure 19.1. Inform

Signatories will proactively inform their employees of an Al Office mailbox where they can submit whistleblower complaints, provided such a mailbox is operational.

OPEN QUESTIONS

- Are there other parts of EU Directive 2019/1937 (the "whistleblowing directive") that are important to highlight in the Code?
- Are there parts of the whistleblowing directive that should be clarified or further specified in the Code? Are there additional whistleblowing measures that may be appropriate to enable assessment and mitigation of 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. Such notifications will be done with understanding of the AI Office's obligations to protect the confidentiality of information provided as per Article 78.

In order to satisfy Measure 20.:

Sub-Measure 20.1. General-purpose AI model with systemic risk notification

LEGAL TEXT

Article 52(1): "Where a general-purpose AI model meets the condition referred to in Article 51(1), point (a), the relevant provider shall notify the Commission without delay and in any event within two weeks

> after that requirement is met or it becomes known that it will be met. That notification shall include the information necessary to demonstrate that the relevant requirement has been met. If the Commission becomes aware of a general-purpose AI model presenting systemic risks of which it has not been notified, it may decide to designate it as a model with systemic risk."

Signatories will, before starting a training run, estimate the amount of computational power they intend to use and notify the AI Office if that classifies the general-purpose AI model as a general-purpose AI with systemic risk.

OPEN QUESTION

The AI Office has the authority to update the classification criteria for determining whether a general-purpose model is presumed to have high-impact capabilities (and therefore whether it is classified as a general-purpose AI model with systemic risk). How could it be written such that it is clear when providers should notify the AI Office of a model meeting new classification criteria?

Sub-Measure 20.2. SSF notification

Signatories will ensure the AI Office has access to the latest version of their Safety and Security Framework.

OPEN QUESTION

How can this access be facilitated?

Sub-Measure 20.3. SSR notification

Signatories will send the AI Office SSRs ahead of the decisions they pertain to, in particular before placing a new general-purpose AI model with systemic risk on the market.

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.

OPEN QUESTION

What constitutes strong reason to believe systemic risk might materialise?

Measure 21. Documentation

Signatories commit to document evidence relevant to their adherence to the Code and the provisions on general-purpose AI models with systemic risk in the AI Act, throughout the lifecycle of the general-purpose AI model with systemic risk, for the purposes of sharing this information with the AI Office upon request.

This includes evidence relevant to the classification of general-purpose AI models with systemic risks, such as the information in Annex XIII. It also includes documentation substantiating their adherence to the AI

Act and the Code, such as SSFs, SSRs, and any additional evidence collected during risk assessments, in addition to the information outlined in Annex XI, Section 2 (see Article 53(1)(a)).

OPEN QUESTION

What could a standardised template for such documentation look like, to reduce compliance costs, especially for smaller providers? Note: in future drafts, we intend to ensure the documentation under this Measure is streamlined and combined other documentation requirements such as those detailed in Annex XI, Section 1, and Annex XII.

Measure 22. Public transparency

Signatories commit to offer appropriate public transparency with the aim of aiding the wider ecosystem, including downstream providers, the AI Office, and the public, to better understand and mitigate systemic risks stemming from general-purpose AI with systemic risk, especially in light of the nascency of the science of assessing and mitigating risks of AI, by publishing their SSF and SSRs. Information may be redacted where its inclusion would substantially increase systemic risk or divulge sensitive commercial information to a degree disproportionate to the societal benefit.

OPEN QUESTIONS

- For what types and levels of public transparency do systemic risks increase, instead of decreasing by empowering the broader ecosystem to assess and mitigate them?
- How burdensome is this kind of public transparency, given the common practice of publishing model and system cards? Can the measure be designed to reduce such burdens?

Conclusion

This first draft of the Code is the result of a preliminary review of existing best practices by the four specialised Working Groups, stakeholder consultation input from nearly 430 submissions, responses from the provider workshop, international approaches (including the G7 Code of Conduct, the Frontier AI Safety Commitments, the Bletchley Declaration, and outputs from relevant government and standard-setting bodies), and, most importantly, the AI Act itself. At this initial stage, the draft is necessarily high-level and primarily sets out principles underlying the Code, together with some proposed Measures and Sub-Measures.

We emphasise that this is only a first draft and consequently the suggestions in the draft Code are **provisional and subject to change**. Therefore, we invite your constructive input as we further develop and update the contents of the Code and work towards a more granular final form for 1 May 2025. Please note in particular that:

1. This first draft is guided by **six key considerations** set out in the drafting plan: i) alignment with Union principles and values, ii) alignment with AI Act and international approaches, iii) proportionality to risks, iv) proportionality to size and capacity of providers, v) support and growth of the AI safety ecosystem, and vi) future-proofing. Together, these principles aim to advance the purposes of the AI Act.
2. We recognise the **need to comprehensively review, develop, and refine** Measures, Sub-Measures, and KPIs based on input from actors with a wide variety of perspectives, including civil society, academia, the AI Safety institutes and industry. Future iterations will be guided by the aforementioned drafting plan and principles, and may include more specific references to articles and recitals of the AI Act. Notwithstanding the preliminary nature of the examples set out above, we welcome detailed comments on Measures, Sub-Measures, and KPIs that future iterations of the Code ought to contain. We also encourage suggestions on how Measures, Sub-Measures, and KPIs can be made more proportionate and more appropriate for different business models and deployment strategies, as well as proposals for how to resolve the open questions we outline within the draft.
3. We note that the current draft is written with the **assumption that there will only be a small number of both general-purpose AI models with systemic risks and providers thereof**. Should that assumption prove wrong, future drafts may need to be changed significantly, for instance, by introducing a more detailed tiered system of measures aiming to focus primarily on those models that provide the largest systemic risks.

Please provide your comments on this draft through our dedicated feedback portal by Thursday, 28 November, 12:00 CET through a form on the dedicated platform (Futurium). We look forward to collaborating with you on future iterations of the Code.