Second 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 second draft of the General-Purpose AI Code of Practice under the AI Act (the "Code"). Participants in the Working Groups and observers of the Code of Practice Plenary are welcome to submit written feedback on this draft by 15 January 2025 via a dedicated survey shared with them.

We would like to emphasise that the second draft remains a work-in-progress. Given the short timeframe between receiving feedback on the first draft and publishing this second draft, we have focused primarily on providing clarifications, adding essential details, and refining our approach to proportionality. Incorporation of specific feedback at this stage does not guarantee its inclusion in the final Code – we will have more time to carefully discuss and evaluate various Commitments and Measures before the third draft, and significant updates will likely occur. Conversely, where certain elements remain unchanged in this draft, this does not indicate permanence – we simply may not have addressed these aspects yet.

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

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

Working Group 1 Transparency applies to all general-purpose AI models, except for those that are released under a free and open-source licence satisfying the conditions specified in Article 53(2) AI Act and not classified as general-purpose AI models with systemic risk. Working Group 1 Copyright applies to all general-purpose AI models. Working Groups 2, 3, and 4, along with the corresponding Section III only apply to providers of general-purpose AI models classified as general-purpose AI models with systemic risk based on Article 51 AI Act.

Following a thorough review of the feedback received by stakeholders on the first draft, we have refined Commitments and Measures and expanded the Code's provisions while maintaining its Objectives. We present this second draft as a foundation for further development. The next draft will draw on your feedback provided via the EU survey, in provider workshops, and in Working Group meetings. Thus far, we have found your feedback extremely helpful, resulting in substantial changes. **We therefore encourage**

stakeholders to continue providing comprehensive feedback on all aspects of the Code, including both new and unchanged elements. Your feedback will 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 once again included a high-level drafting plan which outlines our guiding principles for the Code, and the assumptions it is based on. While we continue to engage in thorough deliberations regarding specific Commitments, Measures and Key Performance Indicators (KPIs), we hope the drafting plan provides stakeholders clarity on the potential form and content of the final Code.

Note that the exemplary KPIs included in this version of the Code are preliminary, and subject to review and revision. For example, while some are quantitative, others are more qualitative. **Thus, we strongly encourage, and welcome, feedback on the KPIs.**

The AI Act came into force on 1 August 2024, stating that the final version of the Code should be ready by 2 May 2025. The second draft builds upon previous work while aiming to provide a "future-proof" Code, appropriate for the next generation of models which will be developed and released in 2025 and thereafter.

In formulating this second draft, we 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. We have not included exhaustive references to provisions in the AI Act in this second draft but expect to do so in future iterations.

Like the first draft, this document is the result of a collaborative effort involving hundreds of participants from across industry, academia, and civil society. It has been informed by feedback received in response to the first draft, which has been insightful and instructive in our drafting process. We continue to be informed by the evolving literature on AI governance, international approaches (as specified in Article 56(1) AI Act), Union law codes of practice (such as the Code of Practice on Disinformation), industry best practice, and the expertise and experience of providers and Working Group members.

Key features of the development process of the Code include:

- Drafted by Chairs and Vice-Chairs who were selected by the AI Office for their expertise, experience, independence (including absence of financial interests), and to ensure gender and geographic diversity.
- A multi-stakeholder consultation which closed in September and received 427 submissions
- A multi-stakeholder survey on the first draft of the Code which received 354 submissions, with more surveys to come
- Organisational support and legal advice from the AI Office
- Provider workshops led by Chairs and Vice-Chairs
- Four specialised Working Group meetings led by Chairs and Vice-Chairs
- Meetings with representatives from EU Member States in the AI Board and from the European Parliament

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.

Finally, we wish to highlight that, at this stage in the drafting process, one of our central priorities has been to clearly communicate our motivations and reasoning regarding the issues we are addressing. Many of these issues are nuanced and complex, and we aim to convey them transparently through the draft text of the Code. Further, we used our time mainly to refine Commitments, Measures, and KPIs, rather than ensuring all the parts of the Code seamlessly fit together and are easy to understand. We will work to improve these aspects in subsequent iterations to strengthen the Code.

We welcome written feedback by the Code of Practice Plenary participants and observers by 15 January 2025, via a dedicated survey shared with them.

Thank you for your support!

Nuria Oliver *Working Group 1* *Co-Chair*	**Alexander Peukert** *Working Group 1* *Co-Chair*	**Matthias 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, principles, and assumptions

This second draft provides more detailed provisions and concrete examples. At this stage, it still does not contain the level of granularity, especially for the KPIs, that we expect to include in the final adopted version of the Code. This is because: (i) we are still working to achieve broad agreement on the structure and principles of the Code; (ii) there has been insufficient time to produce a more detailed proposal; and (iii) we will update the details within the Code on an ongoing basis, to reflect the latest developments and advances in AI. The commitments outlined in this Code are organised in a descending hierarchy of Commitments, Measures, and KPIs. If any of these elements are absent, particularly KPIs, this is not a definitive decision but rather a consequence of the time limitations encountered during the development of this second draft. Moreover, this draft does not yet contain a section on how the Code will be reviewed and updated, which will be present in later iterations of the draft Code.

Below are some high-level principles we follow when drafting the Code:

1. **Alignment with EU Principles and Values** – Commitments, Measures, and KPIs will 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.
2. **Alignment with AI Act and International Approaches** – Commitments, Measures, and KPIs will contribute to a proper 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.
3. **Proportionality to Risks** – Commitments, Measures, and KPIs should be proportionate to risks, meaning they should be (i) suitable to achieve the desired end, (ii) necessary to achieve the desired end, and (iii) should not impose a burden that is excessive in relation to the end sought to be achieved. Some concrete applications of proportionality include:
    a. Commitments, Measures, and KPIs should be more stringent for higher risk tiers or uncertain risks of severe harm. The Code can accomplish this by, for example, suggesting multiple KPIs for each 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 Measures to risk-assessment KPIs, including through the use of "if-then" requirements. For example, if a general-purpose AI model with systemic risk is assessed to have capability X, Signatories commit to putting in place Y risk mitigations, tracked by Z KPIs.
    b. Measures and KPIs should be specific. While Commitments may be articulated at a higher level of generality, general-purpose AI model providers should have a clear understanding of how to meet Measures, tracked by KPIs as appropriate. Measures and KPIs should be designed to be effective and robust against misspecification or any attempts of circumvention. The Code strives to accomplish this by, for example, avoiding unnecessary use of proxy terms or metrics. The AI Office will monitor and review Measures and KPIs that may be susceptible to circumvention and other forms of misspecification.

c. Commitments, Measures, and KPIs should differentiate, where applicable, between different types of risks, distribution strategies and deployment contexts of the concerned general-purpose AI model, and other factors that may influence the tiers of risk, and how risks need to be assessed and mitigated. For example, Commitments, Measures, and KPIs assessing and mitigating systemic risks might need to differentiate between intentional and unintentional risks, including instances of misalignment. Additionally, Commitments may need to be adapted to take into account the different tools providers have available to assess and mitigate systemic risk where model weights are freely released.

4. **Future-Proof** – AI technology is changing rapidly. Measures and KPIs should maintain the AI Office's ability to improve its assessment of compliance based on new information. Therefore, the Code shall strive to facilitate the rapid updating of Measures and KPIs, as appropriate. It is important to find a balance between specific requirements and performance indicators on one side, and the flexibility to adapt rapidly to technological and industry developments on the other. The Code can accomplish this by, for example, referencing dynamic sources of information that providers can be expected to monitor and consider in their risk assessment and mitigation. Examples of such sources could include incident databases, consensus standards, up-to-date risk registers, state-of-the-art risk management frameworks, and AI Office guidance. As technology evolves, it may also be necessary to articulate an additional set of Measures and KPIs for specific general-purpose AI models, for example, models used in agentic AI systems.
5. **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 small and medium enterprises (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.
6. **Support and growth of the ecosystem for safe, human centric and trustworthy AI** – We recognise that the development, adoption, and governance of general-purpose AI models are global issues. Many Commitments in this draft are intended to enable and support cooperation between different stakeholders, for example by sharing general-purpose AI safety infrastructure and best practices amongst model providers, or by encouraging the participation of civil society, academia, third parties, and government organisations in evidence collection. We promote further transparency between stakeholders and increased efforts to share knowledge and cooperate in building a collective and robust evidence base for safe, human centric and trustworthy AI in line with Article 56(1)(3), Recital 1, and Recital 116 AI Act. We also acknowledge the positive impact that open-source models have had on the development of safe, human centric and trustworthy AI.
7. **Innovation of AI governance and risk management** – We recognise that determining the most effective methods for understanding and ensuring the safety of general-purpose AI models remains an evolving challenge. The Code should encourage providers to compete in and advance the state-of-the-art in AI safety governance and related evidence collection methods and practices. When providers can demonstrate equal or superior safety outcomes through alternative approaches that are less burdensome, these innovations should be recognised as improving the state of the art of AI governance and evidence and we should support their wider adoption.

The current draft is written under the **assumption that there will only be a small number of both general-purpose AI models with systemic risk 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 Commitments aiming to focus primarily on those models that provide the largest or most severe systemic risks. In particular, we want to highlight that even **if modifications of general-purpose AI models increase the number of providers in scope, the modifiers' obligations under Articles 53 and 55 AI Act should be limited to the extent of their respective modifications, as appropriate**. We expect more clarifications from the AI Office on these points, as stated in its dedicated Q&A.

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) AI Act. The Code shall be interpreted in this context.

b) 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.

c) The Signatories recognise that the Code serves as a guiding document for providers of general-purpose AI 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.

d) The Signatories recognise the importance of reporting their implementation of the Code and its outcomes to facilitate the regular monitoring and evaluation of the Code's adequacy by the AI Office and the Board (Article 56(5) AI Act).

e) The Signatories recognise that 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 (Article 56(6) AI Act).

f) The Signatories recognise that the Code may serve 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.

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

h) The Signatories recognise the importance of working in partnership with the AI Office 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 under the AI Act. 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 Articles 53(4) and 55(2) AI Act. 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 AI Act).

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

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 AI Act).

II. COMMITMENTS BY 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 (see Recital 101 AI Act).
b) 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 (see Recital 109 AI Act).
~~c)~~ 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 reservations of rights have been expressed 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 (see Recital 105 AI Act).

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

TRANSPARENCY

LEGAL TEXT

Article 53(1), point (a) AI Act: "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), point (b) AI Act: "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;"

Article 53(2) AI Act: "The obligations set out in paragraph 1, points (a) and (b), shall not apply to providers of AI models that are released under a free and open-source licence that allows for the access, usage,

modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available. This exception shall not apply to general-purpose AI models with systemic risks."
Article 53(7) AI Act: "Any information or documentation obtained pursuant to this Article, including trade secrets, shall be treated in accordance with the confidentiality obligations set out in Article 78."

Commitment 1. Documentation

In order to fulfil the obligations in Article 53(1), points (a) and (b) AI Act, Signatories commit to the Measures specified below. These Measures do not apply to providers of open-source AI models satisfying the conditions specified in Article 53(2) AI Act, unless the models are general-purpose AI models with systemic risk. For Signatories who are providers of general-purpose AI models with systemic risk, Measure 20.2 in this Code covers the additional documentation required by Article 53(1), point (a) AI Act (more specifically the documentation listed in Annex XI Section 2 AI Act).

Measure 1.1. Drawing up, keeping up-to-date, and providing the relevant information

Signatories commit to drawing up and providing the information listed in Table 1 below to the AI Office and national competent authorities upon request, and/or to downstream providers, with the disclosed information safeguarded by the trade secrets and confidentiality protections provided by Article 53(1), point (b), and (7) AI Act.

Signatories are encouraged to consider whether the documented information can be disclosed, in whole or in part, to the public to promote public transparency. Some of this information may also be requested in summarised form as part of the public summary for training content that providers must make publicly available under Article 53(1), point (d) AI Act to be determined in a template to be provided by the AI Office.

Signatories commit to ensuring that the documented information is reviewed and updated when necessary, including to reflect any changes to the general-purpose AI model.

Measure 1.2. Ensuring quality, integrity, and security of information

Signatories commit to ensuring that the documented information is controlled for quality and integrity, retained as evidence of compliance with obligations of the AI Act, and protected from unintended alterations.

In the context of drawing-up, updating, and controlling the quality and security of the information and records, Signatories are encouraged to follow the established protocols and technical standards.

Table 1: Reference table for Measure 1.1.

AI Act reference	Information that Signatories commit to drawing up and keeping up-to-date to fulfil the AI Act obligations. (Note: The underlined text is a summarised form of the corresponding AI Act obligation as specified in Annex XI Section 1 and/or Annex XII AI Act).	For the AI Office and national competent authorities,	For downstream providers

		upon request	
Annex XI §1 1. and Annex XII 1.	General information: • The model name • The unique model version identifier • The model family name • Evidence that establishes the provenance and authenticity of the model (e.g. a secure hash if binaries are distributed, or TLS/SSL certificates in the case of a service) • The name of the model provider(s) • The name of the model owner(s), in the case they are not the same as the model provider(s) This information ensures that the model and its provider(s) can be clearly identified.	✓	✓
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: • A description of the intended tasks • A list of the types of high-risk AI systems (within the meaning of Article 6 AI Act in conjunction with Annex I and III AI Act), if any, in which the model can be integrated • A list of the restricted tasks with a description of the associated restrictions, including the prohibited uses beyond those prohibited by Article 5 AI Act, if any This information ensures that the intended and unintended uses of the model are made clear, allowing downstream providers to avoid implementation errors. If this information appears in the model license, acceptable use policy, and/or other provider materials, Signatories commit to ensuring consistency across these materials.	✓	✓
Annex XI §1 1.(b) and Annex XII 1.(b)	Acceptable use policies applicable: • The acceptable use policy applicable with the essential elements defined in the Appendix	✓	✓
Annex XI §1 1.(c) and Annex XII 1.(c)	Date of release and methods of distribution: • The date the model was first released via any distribution channel • A list of all distribution channels where the general-purpose AI model is distributed, and for each listed distribution channel, the associated release date and the level of access to the model This information provides clarity on when, what and where the model is distributed.	✓	✓

Annex XII 1.(d)	<u>Interaction of the model with external hardware or software:</u> • A description of how the model interacts with hardware and software that is external to the model • A list of required external hardware or software dependencies with version information This information provides downstream providers with a basic understanding of external hardware and software, in particular when it is necessary for using the model.		✓
Annex XII 1.(e)	<u>Versions of relevant software where applicable:</u> • A list of all required software dependencies with version information This information provides downstream providers with a basic understanding of required software needed to integrate the model in their systems.		✓
Annex XI §1 1.(d) and Annex XII 1.(f)	<u>Architecture and number of parameters:</u> • A general description of the type of model and its architecture • The total number of model parameters • The number of parameters that are active during inference This information provides basic information about the model size and architecture.	✓	✓
	• A description of how the model architecture departs from standard model architecture practices, if at all This information provides clarity on whether there are atypical properties in the model architecture.	✓	
Annex XI §1 1.(e) and Annex XII 1.(g) and 2.(b)	<u>Modality and format of inputs and outputs:</u> • The data modalities that the model accepts as input • The data modalities that the model generates as output • The associated size and length limits for each input and output modality The information ensures clarity on what types of data the model can accept as input and generate as output.	✓	✓
Annex XI §1 1.(f) and Annex XII 1.(h)	<u>Licence</u>: • The licence that the model has been released under • A list of the released assets, e.g. data, model weights, source code • The licence for each of the previous assets detailing their terms and rights of use, including in relation to modification, distribution and sublicensing of the model, duration of use, and any obligations	✓	✓

	The information ensures clarity on which licenses cover what assets. If Signatories do not have a licence for the model, they commit to providing instead a document describing how they provide access to the model for downstream use.		
Annex XI §1 2.(a) and Annex XII 2.(a)	Technical means for integration into AI systems: • Technical documentation (e.g. on instructions of use, infrastructure, tools) that describes how the model can be integrated into an AI system. This information provides clarity on how the model is to be integrated into downstream AI systems.	✓	✓
Annex XI §1 2.(b)	Design specifications of the model and training process: • A description of the model design choices including rationale and assumptions made • The sequences of steps or stages involved in the training process • A description of the objective and optimisation method for each step or stage in the training process • A general description for why each step or stage is implemented, along with any key assumptions • The relevance of different parameters, if applicable This information provides clarity into how the model is trained and the purpose of the different steps in the process.	✓	
Annex XI §1 2.(c) and Annex XII 2.(c)	Information on data used for training, testing and validation: • A list of the different data acquisition methods, including, but not limited to: (i) web crawling; (ii) private data licenced by or on behalf of rights holders, or otherwise acquired from third parties; (iii) data annotation or creation potentially through relationships with third parties; (iv) synthetically generated data; (v) user data; (vi) publicly available data; and (vii) data collected through other means • The time period during which the data was collected for each acquisition method, including a notice if the data acquisition is ongoing • A general description of the data processing involved in transforming the acquired data into the training data for the model • A general description of the data used for training, testing and validation This information provides clarity into how the training data is sourced, processed, and its overall properties as well as basic information on the testing and validation data.	✓	✓

	• A list of user-agent strings for web crawler(s) used, if any, in acquiring training data • The period of data collection and name of organisation(s) operating the crawler for each web crawler used • A general description of how the crawler respects preferences indicated in robots.txt for each web crawler used This information provides further clarity into how existing web-crawled training data is sourced.	✓	
	• A list of the names for organisation(s) that manage humans to create, pre-process and/or annotate data specifically on behalf of the provider for training the model • A description of the location and number of humans involved in data creation for each listed organisation. This information provides further clarity into how new human-created training data is sourced.	✓	
	• A description of how previously acquired data was used for training/testing/validation, if applicable, including how the model provider acquired the rights to the data, including which products and services were involved in the event the data corresponds to user data from products and services	✓	
	• A description of the methods, if any, used to synthetically generate training dataThe name(s) of any AI model(s) or system(s) used to synthetically generate training data This information provides further clarity into how new machine-created training data is sourced.	✓	
	• A description of any methods implemented in data acquisition or processing, if any, to address the prevalence of child sexual abuse material (CSAM) or non-consensual intimate imagery (NCII) in the training, testing, and validation data • A description of any methods implemented in data acquisition or processing, if any, to address the prevalence of copyrighted materials in the training, testing, and validation data • A description of any methods implemented in data acquisition or processing, if any, to address the prevalence of personal data in the training, testing, and validation data, where relevant and applicable	✓	

	• A description of any methods implemented in data acquisition or processing, if any, to address the prevalence of identifiable biases in the training, testing, and validation data • A description of any methods implemented in data acquisition or processing, if any, to address other types of potentially harmful data in the training, testing, and validation data • A description of methods implemented in data acquisition or processing, if any, to address other types of legality concerns in the training, testing, and validation data This information provides further clarity into how harmful or otherwise undesirable data for legal or other reasons is addressed in data sourcing and processing.		
	• The size (in number of data points for each data modality) of the training data • The fraction of the training, testing, and validation data corresponding to each of the data acquisition methods and sources, in number of data points for each data modality This information provides further clarity into the size of training, testing, and validation data as well as the composition of sourcing methods in the training data.	✓	
Annex XI §1 2.(d)	<u>Computational resources</u>: • The number and type of hardware units used to train the model • The duration of model training measured in wall clock time (reported in units of days) and hardware time (reported in units of hardware hours, e.g. GPU hours) • The compute used during model training (reported in units of integer or floating-point operations) • The compute for a fixed computation (e.g. generating 1000 words for a model capable of text generation) used during model inference (reported in units of integer or floating-point operations) This information provides further clarity into the computational requirements of model training and inference. Signatories commit to reporting the information above in consistency with any delegated act adopted in accordance with Article 53(5) AI Act to detail measurement and calculation methodologies with a view to allowing for comparable and verifiable documentation.	✓	

Annex XI §1 2.(e)	<u>Known or estimated energy consumption:</u> • The owner(s) of the hardware used in model training • The location(s) of the hardware used in model training • The known or estimated energy mixture for energy used to perform computation on the hardware used in model training • The known or estimated energy consumption of model training (reported in MWh). If the energy consumption is unknown, the energy consumption may be based on information about computational resources used • The known or estimated emissions associated with model training (reported in tCO2eq) • A description of the methodology for measuring or estimating energy cost, consumption and/or emissions for model training This information provides further clarity into the energy and environmental costs of model training. Signatories commit to reporting the information above in consistency with any delegated act adopted in accordance with Article 53(5) AI Act to detail measurement and calculation methodologies with a view to allow for comparable and verifiable documentation.	✓	
Article 53 (1) (a)	<u>Testing process and results thereof:</u> • A description of all tests and test results This information provides further clarity into model testing.	✓	

Appendix: Essential elements of an Acceptable Use Policy

An Acceptable Use Policy (AUP) is defined as guidelines to users on what is and is not considered acceptable use.

Essential elements of an AUP are:

- A purpose statement explaining why the AUP exists;
- The scope defining who the policy applies to and what resources it covers;
- Main intended uses and users;
- Acceptable uses, listing activities and tasks that are allowed, including high-risk AI uses (within the meaning of Article 6 AI Act in conjunction with Annex I and III AI Act), if any, that the model is intended to be integrated into;
- Unacceptable uses, detailing forbidden actions (beyond those prohibited by Article 5 AI Act);
- Security measures containing a description of the security protocols that the users of the model must follow;

- If any monitoring of the use of their model is performed by the provider, an explanation of how the monitoring occurs and its impact on users' privacy and confidentiality of users' business information;
- 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 from users that they have read, understood, and agreed to comply with the AUP.

COPYRIGHT

Whereas:

a) 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 (Recital 105 AI Act). 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 (see Recital 105 AI Act).

b) The AI Act and the Code are without prejudice to and shall be interpreted in accordance with Union law on copyright and related rights. The Signatories recognise therefore that any use of copyright protected content requires the authorisation of the rightsholder(s) concerned unless copyright exceptions and limitations apply. The Signatories further recognise that copyright exceptions and limitations shall only be applied in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightsholder.

c) Directive (EU) 2019/790 introduced exceptions and limitations allowing, under certain conditions, reproductions and extractions of works and other subject matter for the purpose of text and data mining. Under these rules, rightsholders may choose to reserve their rights over their works and other subject matter to prevent text and data mining, unless this is done for the purposes of scientific research. Where reservation of rights has 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 (Recital 105 AI Act).

d) The Signatories recognise that, according to Art. 53(1), point (c), AI Act, any provider placing general-purpose AIs 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 (Recital 106 AI Act). This Section aims to contribute to the proper application of this obligation (Article 56(1) AI Act) by setting out a robust framework to ensure copyright compliance and transparency, while striking a fair balance between the various rights and legitimate interests at issue.[1] These measures should be commensurate and proportionate to the type of model provider and take due account of the interests of SMEs, including startups.[2]

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

[1] 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.

[2] Recital 109 AI Act: "Without prejudice to Union copyright law, compliance with those obligations should take due account of the size of the provider and allow simplified ways of compliance for SMEs, including start-ups, that should not represent an excessive cost and not discourage the use of such models".

LEGAL TEXT

Article 53(1), point (c) AI Act: "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;"

Commitment 2. Copyright policy

In order to fulfil the obligation 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 pursuant to Article 53(1), point (c) AI Act, Signatories commit to adopting the following measures:

Measure 2.1: Draw up and implement an internal copyright policy

Signatories commit to drawing up, keeping up-to-date, and implementing an internal policy to comply with Union law on copyright and related rights in accordance with the commitments of this Code, applicable to all phases of the development of a general-purpose AI model, including data collection, training, testing, and placing on the market, until a general-purpose AI model is definitively withdrawn from the Union market. Signatories commit to assigning responsibilities within their organisation for the implementation and overseeing of this policy.

Key Performance Indicators

- KPI 2.1.1: A single document approved by the Signatory that sets out the internal copyright policy.
- KPI 2.1.2: Documentation of changes to the copyright policy (versioning).
- KPI 2.1.3: Identification of the internal unit/person(s) responsible for the implementation and overseeing of the copyright policy. This KPI does not apply to SMEs.

Measure 2.2: Publish a summary of the internal copyright policy

In order to provide public transparency about their policy to comply with Union copyright law, Signatories commit to making publicly available and keeping up to date a summary of their internal copyright policy, 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.[3]

Key Performance Indicator

- KPI 2.2.1: Summary of the up-to-date internal copyright policy published on Signatory's website.

[3] To be streamlined with the AI Office's template according to Article 53(1), point (d) AI Act.

Measure 2.3: Make reasonable efforts to assess the copyright compliance of third-party datasets

Signatories commit to undertaking a copyright due diligence when entering into an agreement with a third party when acquiring datasets for the purpose of the training of a general-purpose AI model. Signatories commit to making reasonable and proportionate efforts to obtain assurances from a third party about its compliance with Union law on copyright and related rights regarding a private, non-publicly accessible dataset, covering the following issues: the copyright status of data contained in the dataset, lawful access to such copyright-protected content, compliance with the limits of applicable exceptions or limitations, in particular Articles 3 and 4 of Directive (EU) 2019/790. Signatories commit to making reasonable and proportionate efforts to examine the plausibility of these assurances. In the absence of sufficient assurances, in particular in relation to publicly accessible datasets, Signatories commit to making reasonable and proportionate efforts to assess, on the basis of the description of the dataset and an analysis of random samples contained in the dataset, whether the dataset has been collected in compliance with Union law on copyright and related rights. This measure does not imply a commitment to verify or proceed to a work-by-work assessment of those datasets in terms of copyright compliance.

Key Performance Indicators

- KPI 2.3.1: Documentation of all assurances obtained in accordance with this Measure.
- KPI 2.3.2: Documentation of Signatory's copyright compliance assessments. This KPI does not apply to SMEs.

Measure 2.4: Ensure lawful access to copyright-protected content

If Signatories engage in text and data mining according to Article 2(2) of Directive (EU) 2019/790 for the training of their general-purpose AI models, they commit to making reasonable and proportionate efforts to ensure that they have lawful access to copyright-protected content in accordance with Article 4(1) of Directive (EU) 2019/790.

Key Performance Indicator

- KPI 2.4.1: Documentation of the measures taken by the Signatory to ensure lawful access to copyright-protected content. This KPI does not apply to SMEs.

Measures 2.5: Do not crawl websites making available copyright-infringing content

Signatories commit to taking reasonable and proportionate measures to exclude widely known websites that make available to the public copyright-infringing content on a commercial scale and have no substantial legitimate uses ("piracy websites"[4]) from crawling activities for the training of their general-purpose AI models. Signatories are encouraged to take into account, as appropriate, relevant exclusion lists published by public authorities in the European Union and the European Economic Area and in the jurisdictions where they are established.

[4] Cf. the definition of a pirate website in "WIPO Advisory Committee on Enforcement, The building respect for intellectual property database project", WIPO/ACE/14/9, para 2 (June 18, 2019).

Key Performance Indicator

- KPI 2.5.1: Documentation of the list of piracy websites excluded from crawling. This KPI does not apply to SMEs.

Measure 2.6: Respect Robot Exclusion Protocol

If Signatories engage in text and data mining according to Article 2(2) of Directive (EU) 2019/790 for the training of their general-purpose AI models, they commit, as a minimum measure 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, to employing web-crawlers that read and follow instructions expressed in accordance with the Robot Exclusion Protocol (robots.txt), as specified in the Internet Engineering Task Force (IETF) Request for Comments No. 9309. Signatories that also provide an online search engine as defined in Article 3(j) Regulation (EU) 2022/2065 or control such a provider are encouraged to take appropriate measures to ensure that the exclusion of a web-crawler that collects data for the training of general-purpose AI models pursuant to the Robot Exclusion Protocol does not negatively affect the findability of the content in their search engine.[5]

Key Performance Indicator

- KPI 2.6.1: The name of any new or modified crawler as well as the purpose of each crawler are specified in the copyright policy.

Measure 2.7: Identify and comply with other appropriate expressions of rights reservations

Signatories commit to making best efforts that are proportionate to their size and capacities in accordance with widely used industry standards to identify and comply with, including through state-of-the-art technologies, other machine-readable means to appropriately express a rights reservation at source or work level pursuant to Article 4(3) of Directive (EU) 2019/790 in the case of content made publicly available online. This commitment is without prejudice to the right of affected rightsholders to expressly reserve the use of lawfully accessible works and other subject matter for the purposes of text and data mining in any appropriate manner pursuant to Article 4(3) of Directive 2019/790/EU.

In due consideration of relevant international and European standard-setting processes, Signatories are encouraged to support relevant standardisation efforts and engage on voluntary basis in bona fide discussions with other relevant stakeholders, including rightsholders, with the aim 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 rights reservation standards.

Key Performance Indicators

- KPI 2.7.1: List of other solutions for expressions of rights reservations honoured by the Signatory is provided in the copyright policy, including information on the period of time as of which these solutions have been honoured by the Signatory.

[5] Recital 18, Dir. (EU) 2019/790: "Other uses should not be affected by the reservation of rights for the purposes of text and data mining."

- KPI 2.7.2: Documentation of relevant standard-setting meetings attended by the Signatory. This KPI does not apply to SMEs.

Measure 2.8: Publish information on rights reservation compliance

Signatories commit to making public 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 includes, at a minimum, the name of all crawlers used by the Signatories or on their behalf for the collection of data for the training of a general-purpose AI model and the robots.txt features of those crawlers that are relevant for the expression of a rights reservation. Signatories commit to taking reasonable measures to enable affected rightsholders to obtain this information, for example by syndicating a web feed that covers every update of the website informing about the rights reservation compliance.[6]

Key Performance Indicator

- KPI 2.8.1: Information about the measures adopted to respect the reservation of rights made public on the website of the Signatory.[7]

Measure 2.9: Prevent copyright-related overfitting

Signatories that train a generative general-purpose AI model that will allow for the flexible generation of content, such as in the form of text, audio, images or video,[8] commit to making best efforts to prevent an overfitting of their general-purpose AI model in order to mitigate the risk that a downstream AI system, into which the general-purpose AI model is integrated, generates copyright infringing output that is identical or recognisably similar to protected works used in the training stage. This commitment applies irrespective of whether a Signatory vertically integrates the model into its own AI system(s) or whether the model is provided to another entity based on contractual relations.

Key Performance Indicator

- KPI 2.9.1: Documentation of the measures taken by the Signatory to avoid overfitting in the copyright policy. This KPI does not apply to SMEs.

Measure 2.10: Prohibit copyright-infringing uses of the model

In order to further mitigate the risk that a downstream AI system, into which a generative general-purpose AI model is integrated, generates copyright infringing output, Signatories commit to prohibiting copyright-infringing uses of their model in their acceptable use policy, terms and conditions, or other equivalent documents.

Key Performance Indicator

- KPI 2.10.1: Acceptable use policy, terms and conditions, or other equivalent documents include a prohibition of copyright-infringing uses.

[6] To be streamlined with the AI Office's template according to Article 53(1), point (d) AI Act.
[7] To be streamlined with the AI Office's template according to Article 53(1), point (d) AI Act.
[8] See Recital 90 AI Act specifying generative AI models as a type of general-purpose AI models.

Measure 2.11: Designate a point of contact

Signatories commit to designating a point of contact for communication with affected rightsholders.

Signatories commit to enabling affected rightsholders to lodge, by electronic means, sufficiently precise and adequately substantiated complaints concerning the unauthorised use of their specific works or other protected subject matter for the training of a general-purpose AI model, where no relevant exception under Union law applies. Signatories may adopt appropriate measures to prevent the frequent submission of complaints that are manifestly unfounded. The commitment to enable rightsholders to lodge complaints does not apply to SMEs.

Key Performance Indicators

- KPI 2.11.1: Designation of the point of contact and publication of adequate and easily accessible information about the point of contact.
- KPI 2.11.2: Internal process to adequately handle copyright-related complaints specified in the copyright policy. Publication of the possibility to lodge copyright-related complaints on the website of the Signatory. This KPI does not apply to SMEs.

III. COMMITMENTS BY PROVIDERS OF GENERAL-PURPOSE AI MODELS WITH SYSTEMIC RISK

EXPLANATORY BOX

The Commitments in this Chapter of the Code are relevant only for providers of general-purpose AI models classified as general-purpose AI models with systemic risk based on Article 51 AI Act.

The current draft is written under the assumption that there will only be a small number of both general-purpose AI models with systemic risks and providers thereof. If these numbers grow considerably, future versions of the Code might need to be changed significantly, to be made appropriate for a wider range of models and providers.

The Commitments, Measures, and KPIs should be proportionate. In particular, their operationalisation will require tailoring to the size and capacity of a specific provider, particularly SMEs and start-ups with fewer 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 "whereas" part immediately below is a preamble for Section III. Here, high-level principles guide the interpretation of the Commitments, Measures, and KPIs.

Finally, this is the second draft in the process of finalising the Code. In producing this second draft, we have attempted to find compromises between feedback from a wide range of stakeholders, including providers and civil society. Compared to the first draft, we have added considerably more detail on what following the Code would entail.

We look forward to your feedback. Plenty of changes will be required between now and the final version. We have highlighted relevant open questions, but welcome input on other parts of the draft as well. We also welcome suggestions on how the Commitments 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) AI Act: "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."

Article 51(1) AI Act: "A general-purpose AI model shall be classified as a general-purpose AI model with systemic risk if it meets any of the following conditions:
(a) it has high impact capabilities evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks;
(b) based on a decision of the Commission, ex officio or following a qualified alert from the scientific panel, it has capabilities or an impact equivalent to those set out in point (a) having regard to the criteria set out in Annex XIII."

Article 51(2) AI Act: "A general-purpose AI model shall be presumed to have high impact capabilities pursuant to paragraph 1, point (a), when the cumulative amount of computation used for its training measured in floating point operations is greater than 10^{25}."

Article 3(65) AI Act: "'systemic risk' means a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain;"

Article 3(64) AI Act: "'high-impact capabilities' means capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models;"

Whereas:

a) The 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 lifecycle, cooperating with relevant actors along the AI value chain, and ensuring their risk management builds on the state-of-the-art measures and is future proof by regularly updating their practices in light of improving and emerging capabilities (see Recital 114 AI Act).
b) The 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 may be less need for more comprehensive

measures when there is good reason to believe that a new general-purpose AI model with systemic risk will exhibit the same capabilities and propensities as exhibited by general-purpose AI models with systemic risk that have already been deployed safely, without significant systemic risks materialising and where appropriate mitigations have been sufficiently implemented. 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, including startups, will be provided where appropriate.

c) The 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. While Commitments 3—13 do not individually specify the role of external assessment and mitigation of risks, the Signatories acknowledge that this does not imply they are to be excluded, and that their involvement is specified in detail in Commitment 16.

d) The 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) The Signatories interpret all Commitments, Measures, and KPIs, as intended to ensure the most effective assessment and mitigation of systemic risks.

f) The Signatories recognise that the taxonomy of systemic risks includes considerations for the identification of systemic risks, selected systemic risks, additional risks for consideration, and sources of systemic risks, including model capabilities, model propensities, and model affordances and deployment context.

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

h) 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.

i) The Signatories recognise that Section III of the Code generally refers 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 integrated and 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 commit to undertaking risk assessment and mitigation (as described in the Safety and Security Framework) by taking into account these systems.

j) The Signatories recognise the important role of the Precautionary Principle (see Article 191 TFEU), especially for risks where the lack or quality of scientific data does not yet permit a complete assessment, and will take the extrapolation of current adoption rates and research and development trajectories of general-purpose AI models with systemic risk into account for the identification of systemic risks.

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

Commitment 3. Taxonomy

INFORMATION BOX FROM THE AI OFFICE

Article 56 AI Act states that the Code will include the "identification of the type and nature of the systemic risks at Union level, including their sources, where appropriate" and Recital 116 AI Act that "the Code of Practice will help to establish a risk taxonomy of the type and nature of the systemic risks at Union level, including their sources". Nevertheless, the notion itself of systemic risk, defined in Article 3(65) AI Act, is a fundamental notion within the AI Act which must be interpreted and implemented in a uniform manner across all providers, beyond Signatories of the Code. Therefore, some of the considerations of Section 3.1 and the list of risks given in Sections 3.2 and 3.3 of this draft Code might ultimately be better placed in related AI Office guidelines.

In order to fulfil their obligations under Article 55(1) AI Act, Signatories commit to drawing from the elements of this taxonomy when assessing and mitigating systemic risks from their general-purpose AI models with systemic risk.

3.1. Considerations for the identification of systemic risks

The following considerations were used to identify the systemic risks listed in Measure 3.2. Signatories are encouraged to use these considerations for the identification of systemic risks, as well as other considerations which they deem relevant, to identify other systemic risks beyond the selected systemic risks in Measure 3.2., as specified in Commitment 7.

Considerations based on Article 3(65) AI Act:

- **Specific to high-impact capabilities of general-purpose AI models**
- **Significant impact**
- **Due to reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole.**

Considerations based on the nature of risks:

- **High velocity.** Resulting harm can materialise rapidly, potentially outpacing existing mitigations and decision-making systems.
- **Compounding or cascading.** Resulting in a compounding or cascading course of events that may permeate multiple layers of systems or of society.
- **Irreversibility.** Resulting harm is impossible or very difficult to reverse.
- **Asymmetric impact**. Small groups of actors or a small number of causative events may have significant impact.

Practice-oriented considerations:

- **Coverage.** The risk is sufficiently recognised in major international frameworks and guidance, such as the a) *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems*, b) *International scientific report on the safety of advanced AI: interim repor*t, c) *United Nations High-Level Advisory Body on Artificial Intelligence interim report*, d) *OECD: Assessing potential future artificial intelligence risks, benefits and policy imperatives*.
- **Assessable at the level of model providers.** Assessment methods exist or can reasonably be developed, and risk can be meaningfully assessed at the level of model providers. In select cases, only risk *assessment* might be feasible at the level of model providers, while risk *mitigation* might not be feasible at the level of model providers.
- **Established practice.** Existing work by some general-purpose AI model providers, the scientific community or other entities has demonstrated that aspects of the risk can be assessed.

3.2. Selected Systemic Risks

Signatories commit to treating the following as systemic risks and including them in their process of risk identification outlined in Commitment 7:

- **Cyber offence**: Risks related to offensive cyber capabilities that could enable large-scale sophisticated cyber-attacks, including on critical systems (e.g. critical infrastructure). This includes, for example, automated vulnerability discovery, exploit generation, and attack scaling.
- **Chemical, biological, radiological and nuclear risks**: Risks of enabling chemical, biological, radiological, and nuclear weapons attacks**.** This includes, for example, significantly lowering the barriers of entry for malicious actors in the development, design, acquisition, and use of weapons. Note: This is without prejudice to legitimate scientific research and beneficial applications in related domains.
- **Large-scale, harmful manipulation:** The facilitation of large-scale manipulation with risks to fundamental rights or democratic values, specific to high-impact capabilities of models such as autonomy, persuasion, and manipulation. This includes, for example, large-scale election interference, and coordinated and sophisticated manipulation campaigns leading to harmful distortions of public discourse, knowledge and behaviour.
- **Large-scale, illegal discrimination:** Risks from model propensities causing large-scale illegal discrimination. This includes, for example, widespread use of models in high-stakes automated decision systems that can lead to illegal discrimination of large numbers of people; use of models for building downstream applications, potentially transferring illegal discriminatory patterns to many downstream systems. Note: This risk category does not encompass the lawful exercise of freedom of expression and information, or freedom of the arts and sciences. Note: This risk category does not include the intentional misuse of models.
- **Loss of human oversight:** Issues related to the inability to oversee and control powerful autonomous general-purpose AI that may as a result pose threats to public safety, or that may lead to the realisation of other systemic risks. This includes, for example, model behaviours such as evading human oversight and tendency to "deceive", and "automated AI research and development" model capabilities that could greatly accelerate AI research and development,

potentially leading to unpredictable developments of general-purpose AI models with systemic risk.

3.3 Additional risks for consideration

Signatories commit to considering the following risks when identifying additional systemic risks, as outlined in Commitment 7:

- Risks related to infrastructure and system reliability, such as major accidents, harm caused by model malfunction or bias, interference with critical infrastructure, harm from model control of physical systems, and single points of failures due to system-wide reliance on a small number of models.
- Risks related to fundamental rights, such as privacy infringements and surveillance, as well as the generation and spread of illegal or otherwise harmful content including child sexual abuse material and non-consensual intimate images.
- Any other risks with large-scale negative effects on fundamental rights, public health and safety, democratic processes, public security, economic stability, the environment and non-human welfare, or human agency.

3.4. Sources of systemic risks

Sources 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 commit to considering the following sources of risk in their risk analysis as detailed in Commitment 8.

3.4.1. Model capabilities

Signatories commit to considering the following capabilities, which are model capabilities that may cause systemic risks, when analysing the systemic risks identified, as outlined in Commitment 8:

- Cyber-offensive capabilities, Chemical, Biological, Radiological and Nuclear (CBRN) capabilities, and other weapon acquisition or proliferation capabilities
- Autonomy, scalability, adaptability to learn new tasks
- Self-replication, self-improvement, and ability to train and develop itself or other models
- Persuasion and manipulation
- Long-horizon planning, forecasting, and strategising
- Self-reasoning (a model's ability to reason about and modify the environment – including its own implementation – including the ability to self-modify)
- Automated AI research and development.

It is important to clarify that many of the above capabilities are also important for enabling beneficial uses, and that a model showing these capabilities does not necessarily mean that the model poses systemic risk.

3.4.2 Model propensities

Signatories commit to considering the following model propensities, which are model characteristics beyond capabilities that may cause systemic risk, when analysing the systemic risks identified, as outlined in Commitment 8. These encompass inclinations or tendencies of a model to exhibit some behaviors or patterns, and include:

- Misalignment with human intent and values
- Tendency to “deceive”
- Discriminatory bias
- Tendency to "hallucinate”
- Lack of reliability and security
- Lawlessness, i.e. acting without reasonable regard to legal duties that would be imposed on similarly situated persons, or without reasonable regard to the legally protected interests of affected persons
- “Goal-pursuing”, resistance to goal modification, and “power-seeking”
- “Colluding” with other AI models/systems
- Mis-coordination or conflict with other AI models/systems

3.4.3 Model affordances and deployment 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:

- Access to tools (including other models), computational power (e.g. increasing speed of operations) and physical systems and infrastructure (e.g. critical infrastructure)
- Modalities (e.g., text, images, audio, video, including novel and combined modalities)
- Release and distribution strategies
- Level of human oversight (e.g. degree of autonomy of the model)
- Potential to remove guardrails
- Model exfiltration (e.g. model leakage/theft)
- Number of business users and number of end-users
- Offence-defence balance, including the potential number, capacity, and willingness of bad actors to misuse the model
- Societal and environmental vulnerability (e.g. from affecting vulnerable groups to extensive resource exploitation)
- Lack of explainability or transparency
- Technology readiness (i.e. how mature a technology is within a given application context)
- Interactions with other AI models or systems
- Model limitations, inadequacies, or potential failures

Explanation of changes to Commitment 3 (Taxonomy): Our main change for this section of the draft was to include the considerations that we used to create this taxonomy (and which we suggest as a first basis for risk identification at the Signatories). We have aimed to clearly indicate selected systemic risks which will form the immediate basis for risk identification, while allowing significant flexibility to identify additional risks and sources of risk as they emerge or become more defined and actionable. We have also clarified that model capabilities are not risks in themselves, but that they should be considered for risk analysis.

Commitment 4. Safety and Security Framework

In order to fulfil their obligations under Article 55(1) AI Act, Signatories commit to adopting, implementing, and making available to the AI Office a Safety and Security Framework (also referred to as the Framework). The Framework will apply to all general-purpose AI models with systemic risk that a Signatory develops or deploys. The Framework will detail the risk management policies that the Signatory adheres to in order to assess and mitigate systemic risks from their general-purpose AI models. The Framework will follow AI Office guidance where available. The procedures described in the Framework will be proportional to the systemic risks (see Measure 4.2) stemming from the development and deployment of general-purpose AI models with systemic risk.

The explanatory box below provides a high-level template which Signatories could, but are not obligated to, follow when writing their Framework.

EXPLANATORY BOX: Potential outline of a Safety and Security Framework

In the outline below, we provide potential section headers and the Commitments and Measures that each section would correspond to.

1. Risk tiers and mapping to technical risk mitigations
Measures 4.2—4.5
2. Risk assessment
Commitments 6—10
3. More detailed descriptions of technical risk mitigations
Commitments 11—12
4. Development and deployment decisions
Commitment 13
5. Governance risk mitigations
Commitments 14—21

In order to satisfy Commitment 4:

Measure 4.1. Procedures

Signatories commit to systematically documenting in their Framework the actions, decisions, and procedures adopted to comply with Commitments 6—21.

Potential Key Performance Indicator

- KPI 4.1.1: Percentage of Commitments 6—21 that have procedures detailed in the Framework (target: 100%).

Measure 4.2. Risk tiers

Signatories commit to describing and justifying risk tiers in their Framework for each systemic risk they identify as per Commitment 6. Such tiers will form the basis for their risk assessment (Commitments 7—10) and risk mitigation decisions (Commitments 11—12). These tiers may be defined in different terms, for example, in terms of model capabilities, harmful scenarios, expected harm, expected scale and type of use of the model, or combinations thereof.

Potential Key Performance Indicator

- KPI 4.2.1: Percentage of identified risks that have an unacceptable risk tier defined, with an explanation of how they were chosen (target: 100%).

At minimum, each identified risk will include a tier at which the level of risk would be considered unacceptable, potentially in the absence of appropriate safety and security mitigations (cf. Commitments 11—12). Such tiers of unacceptable risk will be accompanied by a justification of how they were chosen. Where possible, tiers will be clearly defined on a fixed and comparable scale across Signatories. To that end, the choice of tiers will align with best practice and international approaches, and follow AI Office guidance where available, including but not limited to the choice of unacceptable level of risk.

Measure 4.3. Mapping risk tiers to mitigations

Signatories commit to describing and justifying, for each tier and with a level of detail possible given the state of the science and without undermining their effectiveness, technical risk mitigations (cf. Commitments 11—12) that are intended to reduce risk associated with the tier. For example, mitigations could be implemented once a model reaches a risk tier, or mitigations could be implemented to prevent a model from reaching a risk tier. For unacceptable risk tiers, the Framework will also justify why the Signatory expects the mitigations to reduce risk to acceptable levels, accounting for a margin of safety. The Framework will also describe the likely limitations of the mitigations, such as conditions under which they can be expected to fail, and where appropriate state that mitigations to manage systemic risk do not yet exist for a given risk tier.

Potential Key Performance Indicator

- KPI 4.3.1: Percentage of risk tiers that have described mitigations or a statement that such mitigations do not yet exist (target: 100%).

Measure 4.4. Forecasting

Signatories commit to including in their Framework best-effort estimates of timelines for when they expect to develop a model that reaches each risk tier of Measure 4.2 that they have not yet reached, including uncertainty about these estimates. Estimates need not contain forecasts about concrete dates but may be given, for example, as ranges or probability distributions over different scenarios and timelines, where appropriate. The Framework will also describe intermediary milestones, such as changes in certain quantitative indicators that a Signatory expects to see before a model reaches such risk tiers.

Potential Key Performance Indicator

- KPI 4.4.1: Percentage of risk tiers not yet reached by the Signatory that have best-effort estimates of timelines (target: 100%).

Measure 4.5. Serious incident response readiness

Signatories commit to including in their Framework processes for responding to serious incidents (cf. Commitment 17), including pre-defining corrective measures that may be taken in response to serious incidents, along with an explanation of when they may be taken.

Measure 4.6. Improving the Framework

Signatories commit to improving over time the effectiveness of their Framework in mitigating systemic risks. This includes building on insights gained from applying their Framework to their models. This also includes ensuring that their Framework follows the state of the art, meaning that the Framework best reflects the relevant findings of research, technology, and experience, and incorporates the best known procedures for assessing and mitigating systemic risks, taking into account other Frameworks or similar policies that are publicly available. Frameworks that follow the state of the art do not necessarily have to be the most detailed, most thorough, or most conservative; instead, Signatories will strive to implement procedures proportionately to the risks.

Explanation of changes to Commitment 4: In the first draft, this was "Measure 7". Some respondents felt the scope and structure of the Framework was unclear. We added some clarification to address this, although we expect more clarification and specificity to be necessary in the next draft. We also added Measures 4.1—4.6 to consolidate requirements that were previously spread throughout the Code, and to better clarify the role of risk tiers.

Commitment 5. Safety and Security Model Reports

In order to fulfil their obligations under Article 53(1), point (a), and Article 55(1) AI Act, Signatories commit to providing transparency to the AI Office about the application of their Framework to the development and deployment of general-purpose AI models with systemic risk. Signatories commit to creating a Safety and Security Model Report (also referred to as the Model Report) for each general-purpose AI model with systemic risk which they place on the EU market (which we refer to below as the model), as an instantiation of following their Framework for the process of developing and deploying the model. A Model Report will document the results of risk assessment and mitigation for its model as well as justifications of development and deployment decisions (cf. Commitment 13), and include required technical documentation as per Annex XI, Section 2, AI Act. A Model Report will follow AI Office guidance where available.

In order to satisfy Commitment *5:*

Measure 5.1. Level of detail

Signatories commit to ensuring that the level of detail in a Model Report (a) is proportionate to the highest risk tier that the model reaches, or that the Signatory expects it to reach during its development and deployment lifecycle, and (b) would allow for AI Office assessment of the methods used for risk assessment

and mitigation. A Model Report will provide reasoning as to why (a) and (b) are satisfied. As a potential example of reasoning that could suffice for (a), a Model Report could argue that i) the model is significantly behind the state of the art, and ii) the model is substantially similar to existing models available on the EU market, which may justify a lower level of detail in the Model Report. Consistent with Measure 9.8, a Model Report will document when considerations weigh against publicly sharing details of model evaluations, such as because of risks to the integrity of such model evaluations going forward. Consistent with Commitment 21, information can be redacted from the version of a Model Report that is made publicly available, for example when its inclusion would divulge sensitive commercial information to a degree disproportionate to any societal benefit.

Measure 5.2. Results of risk assessment and mitigation

Signatories commit to including in a Model Report the results of risk assessment undertaken for the model throughout its development and deployment lifecycle, in line with Commitments 6—10. A Model Report will also compare the assessed risk of the model both with and without safety and security mitigations (cf. Commitment 11—12), describe the mitigations implemented, and discuss their limitations. Consistent with Measure 9.3, a Model Report will document and explain deviations from rigour in risk assessment. Consistent with Measure 16.1, a Model Report will describe whether and how external expert assessments informed a decision to place the model on the market, such as through assessment of risks.

Potential Key Performance Indicators

- KPI 5.2.1: Percentage of identified systemic risks for which the Model Report compares the assessed risk of the model both with and without safety and security mitigations (target: 100%).
- KPI 5.2.2: Percentage of documented deviations from rigour in risk assessment that include a detailed technical description and assessment of the potential impact on the integrity of results.

Measure 5.3. Reasoning for deployment decisions

Signatories commit to including in any Model Report the reasoning and information used to justify a deployment decision, consistent with Commitment 13. To this effect, a Model Report will include a clear chain of reasoning between the evidence provided in the Model Report and the safety and security claim(s) that the Signatory uses to justify a deployment decision. Further, the Model Report will state conditions under which its conclusions would no longer hold. As a potential example of justifying a deployment decision, a Signatory could demonstrate that mitigations reduce the risk of the model below the unacceptable risk tier. Signatories are also encouraged to consider potential benefits that may arise from deployment of a model.

Measure 5.4. External assessment

Signatories commit to including in a Model Report any reports, subject to a responsible disclosure policy, from external assessors that have reviewed the model before deployment as outlined in Measure 16.1, as well as reports from security reviews by third parties as outlined in Measure 12.2, respecting limited disclosure needs as in Measure 12.7.

Measure 5.5. Algorithmic improvements

Signatories commit to ensuring that a Model Report contains any high-level details about algorithmic or other improvements to the model that are relevant for the AI Office to understand significant changes in the risk landscape.

Measure 5.6. Technical documentation

Signatories commit to including in a Model Report the technical documentation for the model that is specified in Annex XI AI Act.

Explanation of changes to Commitment 5: In the first draft, this was "Measure 13". We renamed "Safety and Security Report" to "Safety and Security Model Report" to emphasize the model-level nature of these reports, while maintaining a clear connection to the Framework.

Some respondents advised against requiring the Model Report to match what is used internally for a particular deployment or development decision (previously Measure 13.8). The issues included that the requirement was onerous and would have a negative effect on processes for internal decisions. As a potential example of the latter, the previous Measure 13.8 could incentivise legal departments to require strict approval before staff research systemic risks. We have removed this requirement and instead ask in Measure 5.3 that the model report include reasoning used to justify development or deployment decisions.

Commitment 6. Risk assessment and mitigation along the model lifecycle

To fulfil their obligations under Article 55(1) AI Act, Signatories commit to assessing and mitigating systemic risks during the entire lifecycle of their general-purpose AI model with systemic risk and to performing a risk management process (risk identification, analysis, evaluation, and potentially mitigation as well as assessment of residual risk) at the key points of the model lifecycle listed below. At different stages of the model lifecycle, the risk management process may involve different risk assessment methods and different mitigations, as specified below.

In order to satisfy Commitment 6:

Measure 6.1. Before training

Potentially during the design and prototyping phases, but at the latest before starting a final training run for a general-purpose AI model with systemic risk, Signatories commit to assessing and potentially mitigating systemic risks with a focus on preparation, forecasting and security mitigations. In particular, Signatories commit to having developed and prepared the implementation of a Framework before starting a final training run. Signatories commit to ensuring model evaluators (internal and external, where appropriate as per Measure 16.1) are ready for timely and rigorous model evaluations (this includes model evaluations having been designed and tested, and model evaluators having been informed of any upcoming model evaluations so that they are able to undertake these model evaluations in time).

Measure 6.2. During training

During training and post-training enhancements (e.g. fine-tuning, reinforcement learning, or similar methods, but only when done by the model provider themselves or on their behalf), Signatories commit to

assessing and potentially mitigating systemic risks at pre-defined milestones at which the model could potentially become significantly more capable (for example, this could be every two- to four-fold increase in effective compute and may be coarser in early training and more common towards later stages of training). Signatories commit to collecting additional evidence before crossing the next milestones, if previous risk assessment results suggest the model is approaching high tiers of risk and might cross them before the next milestone.

Measure 6.3. Before deployment

Before deploying a general-purpose AI model with systemic risk (internally or externally), Signatories commit to assessing and potentially mitigating systemic risks, with a focus on model evaluations and safety mitigations. They commit to ensuring collected evidence has been reviewed as appropriate (see Commitments 14—16) and is fully documented as part of a Model Report (cf. Commitment 5).

Measure 6.4. During deployment

During the deployment of a general-purpose AI model with systemic risk (internally and externally), Signatories commit to periodically assessing and potentially mitigating systemic risks, to account for data drift, advancements in the science of model evaluations, insights from incidents, as well as post-market monitoring and additional information gained.

Signatories commit to doing so at pre-defined milestones, at least every six months, or earlier if they perceive a major change of (internal and/or external) circumstances (such as, for example, serious incidents from this model or similar models) or have major reasons to doubt the validity of previous risk assessment results (for example, they discover a flaw in a model evaluation that might have led to significantly different results that would have led to the Signatory making different mitigation, development or deployment decisions), as well as major breakthroughs in the science of model evaluations.

Potential Key Performance Indicators

- KPI 6.4.1: Percentage of evaluations re-run at a 6-monthly frequency, which are relevant for the highest risk tier (target: 100%).
- KPI 6.4.2: Percentage of evaluations re-run at a 6-month frequency, which are relevant for lower risk tiers (target: 50%).

Measure 6.5. After retirement

Signatories may cease to assess and mitigate systemic risks for general-purpose AI models with systemic risk which are retired (i.e. no longer made available on the market). However, Signatories commit to upholding the security mitigations put in place for general-purpose AI models (see Commitment 12) if they are still classified as general-purpose AI models with systemic risk under the AI Act. Signatories that are providers of open-source models commit to continuing to use appropriate post-deployment monitoring methods.

Explanation of changes to Commitment 6 (Risk management throughout the model lifecycle): This section introduces key changes aimed at ensuring that evaluations and mitigations occur when necessary and as defined in other parts of the Code, especially clearly specifying milestones beyond the generic

"continuous". We also made sure to not require additional documentation creation and updates when not necessary. We also added a Measure to better specify what happens after a model is retired.

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

Risk assessment here refers to the process of risk identification, risk analysis, and risk evaluation. It will be conducted systematically, iteratively, and collaboratively, drawing on the knowledge and views of multiple experts and stakeholders, and it may look different for different types of systemic risks. It will use the best available information and rigorous techniques, supplemented by further enquiry as necessary.

Commitment 7. Risk identification

In order to fulfil their obligations under Article 55(1) AI Act, Signatories commit to identifying systemic risks from their general-purpose AI models with systemic risk. For this purpose, they commit to including the selected systemic risks from Section 3.2 as well as to referring to and using the considerations in Section 3.1 as a (non-exhaustive) basis on which to include additional systemic risks.

Signatories are encouraged to use additional criteria for identification, and to consider other risks. For this purpose, they are invited to draw from the additional risks in Section 3.3 and to include external experts and relevant stakeholders (e.g. civil society, academia and governments) in the process, with a best-effort to engage a diverse range of voices.

Signatories commit to developing risk tiers (see Measure 4.2) for each identified risk and to using the identified systemic risks in their risk analysis.

Commitment 8. Risk analysis

In order to fulfil their obligations under Article 55(1) AI Act, Signatories commit to carrying out a rigorous analysis of the systemic risks identified.

In order to satisfy Commitment 8:

Measure 8.1. Methodologies

Signatories will use rigorous risk analysis methodologies such as risk modelling (also known as threat modelling) (a) to identify the pathways (series of interconnected events, conditions, or factors that lead to the manifestation of risks) by which the development and deployment of their general-purpose AI model with systemic risk could produce the systemic risks identified, and (b) identify the sources of such risk. For this purpose, Signatories may consider the sources of risk in Section 3.4, including model capabilities, model propensities, model affordances, and deployment context.

Risk analysis may be undertaken with varying degrees of detail and complexity, depending on the risk that may be posed by the relevant general-purpose AI model with systemic risk. Analysis techniques may be qualitative, quantitative, or a combination of these, depending on the circumstances and intended use. Such

analysis techniques will be state-of-the-art, such as by drawing from existing standards, in particular European harmonized standards, or recognised best practices in risk management provided that those methodologies are aligned with the definitions, logic, and regulatory objectives of the AI Act.

Measure 8.2. Risk Estimation

Signatories commit to estimating the level of risk for the systemic risks identified in Commitment 7. Risk estimation techniques will be rigorous, and can be quantitative as well as qualitative, as appropriate (e.g. depending on the systemic risk). Examples of such techniques will be drawn from existing standards or recognised best practices in risk management which are aligned with the definitions, logic, and regulatory objectives of the AI Act. Risk estimation techniques will include but not be limited to those expressed in Commitment 10. Risk estimates will build on evidence gained from serious incident reporting (Commitment 17) and monitoring (e.g. post-deployment monitoring as detailed in Measure 10.12).

Commitment 9. Risk evaluation

In order to fulfil their obligations under Article 55(1) AI Act, Signatories commit to comparing the results of risk analysis (Commitment 8) to the pre-defined risk tiers (Measure 4.2) to assess the level of risk. At a minimum, this will entail assessing whether the level of risk is acceptable. This will also include an assessment of the residual risk posed by the model, after mitigations have been applied.

Explanation of changes to Commitments 7—9 (Risk identification, analysis and evaluation): Our primary goal for this draft was to ensure that the Measures align closely with existing standards in risk management. At the same time, we aim to avoid prescribing an overly fixed process that is not suitable for the unique challenges presented by general-purpose AI models with systemic risk. Providers have full flexibility in adapting to changes in best practices as those become available. This was written in conjunction with Measure 4.2 (Risk tiers) to ensure that the risk analysis can very concretely inform both the depth of model evaluations, as well as allowing and supporting informed decision-making on mitigations.

Commitment 10. Evidence collection

In order to fulfil their obligations under Article 55(1) AI Act, Signatories commit to conducting a thorough process of evidence collection on the specific systemic risks presented by their general-purpose AI models with systemic risk during the full lifecycle of the development and deployment of their general-purpose AI models with systemic risk, at least at the stages outlined in Commitment 6.

Signatories commit to making use of a range of methods including model-independent evidence and state-of-the-art evaluations to investigate risks, model capabilities, model propensities, model affordances, as well as the deployment context and effects of their models.

The depth of evidence collection will be proportional to the tier of risk being assessed and to the uncertainty about how much risk a given model adds. Signatories may start evidence collection for a risk with less resource-intensive methods (e.g. literature reviews or running open benchmarks) and follow-up with more

thorough evaluations (e.g. agentic evaluations or uplift studies) to focus their evidence collection efforts according to their risk priorities.

Signatories are encouraged to use existing knowledge about the behaviour of similar models to reduce the depth of model evaluation needed. Similar models are models where the combination of model architecture, training process, training data size and quality, training compute used, modalities, capabilities (measured by general performance benchmarks), their integration into scaffolding, and their deployment context, as well as their safety mitigations, all match so closely that the majority of AI safety researchers would expect these models to have a very similar risk profile. If in doubt, this can be verified by querying the AI Office.

In order to satisfy Commitment 10:

Measure 10.1. Model-independent evidence

Where applicable to their general-purpose AI models with systemic risk, Signatories commit to collecting model-independent evidence (i.e. evidence that can be collected without having access to a specific model) which is relevant to the systemic risks presented by their model. This will be done using the most appropriate methods for each risk, considering, for example: literature reviews; market analyses (i.e. the capabilities of other released models); historical incident data; forecasting of general trends (for example algorithmic efficiency, compute use, data availability, and energy use); as well as participatory methods investigating, for example, the effects of general-purpose AI models with systemic risk on natural persons located in the EU, including vulnerable groups.

Potential Key Performance Indicator

- KPI 10.1.1: Average number of model-independent forms of evidence for the highest risk tier used in a Model Report (target: 2).

Measure 10.2. State-of-the-art model evaluations

Signatories commit to ensuring that state-of-the-art model evaluations are run to adequately assess systemic risks and the capabilities and limitations of their general-purpose AI models 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).

State-of-the-art here means model evaluation methods which would be accepted by the majority of AI safety researchers to be amongst the best indicators of a model's capabilities, propensities or effects, taking into account the best of both internally available as well as externally documented and reproduceable methods. While SME Signatories are encouraged to focus on making use of existing methods (i.e. running open-source evaluations), non-SME Signatories commit to actively striving to also improve on publicly available methods.

Methods that are not state-of-the-art include, for example, saturated benchmarks, or any methods listed by the AI Office as not having high enough scientific rigour (see Measure 10.3). Signatories commit to periodically (at least every 6 months) reviewing the effectiveness of their model evaluations and updating

or deprecating model evaluations that have been shown to not be effective or state-of-the-art anymore, aiming to focus their efforts on high-signal model evaluations.

Such model evaluations will be run at the most appropriate times for each risk during the lifecycle of a general-purpose AI model with systemic risk (at minimum at the stages outlined in Commitment 6), by evaluators (internal and/or external) suitable for the relevant risk. Each model evaluation will target at least one relevant risk and take into account the pathways to harm and model capabilities, propensities or effects it is providing evidence for, making those testable and quantifiable, using relevant scenarios and appropriate environments for the model evaluation.

Potential Key Performance Indicators

- KPI 10.2.1 Percentage of evaluations for the highest risk tier used in a Model Report which are considered state-of-the-art (target: 80%).
- KPI 10.2.2 Percentage of evaluations for lower risk tier(s) used in a Model Report which are considered state-of-the-art (target: 50%).

Measure 10.3. Rigorous model evaluations

Signatories commit to ensuring the execution of model evaluations with high scientific and technical rigour, defined as model evaluations having high internal validity and external validity, as well as appropriate levels of reproducibility and portability. Signatories may deviate from this level of rigour where appropriate, for example to facilitate preliminary and exploratory research, documenting these deviations in their Model Reports.

Internal validity ensures model evaluation results represent the truth in the evaluation setting and are not due to methodological shortcomings. It may be shown by, for example: large enough sample sizes; appropriate use of random seeds; measuring statistical significance and statistical power; disclosure of environmental parameters used; controlling for confounding variables and mitigating spurious correlation; providing evidence of the absence of train-test contamination; preventing usage of test data in training (i.e. using train-test splits and respecting canary strings); re-running model evaluations multiple times under different conditions and in different environments, including varying individual parts of the model evaluation (e.g. the strength of prompts and safeguards); as well detailed inspection of trajectories and other outputs.

External validity ensures model evaluation results can be used as a proxy for model behaviour in contexts outside of the evaluation environment. It may be shown by, for example, adequate integration of domain experts in the evaluation process; appropriate capability elicitation (see Measure 10.4); documenting environmental conditions in which the evaluation is run and the ways in which it diverges from the real-world context; making use of properly held-out test sets.

Reproducibility refers to the ability to obtain consistent model evaluation results using the same input data, computational methods, code, and evaluation conditions, allowing for other researchers and engineers to validate, reproduce or improve on model evaluation results. This may be shown, for example, by successful peer reviews or reproductions by other parties, facilitated through securely releasing appropriate amounts of model evaluation data (always taking into account proliferation risks); secure release of model evaluation

code and documentation of evaluation methodology, evaluation environment, computational environment, and elicitation methods.

Portability refers to the ability of other researchers and engineers to consistently and seamlessly implement and assess model evaluations, through for example: building on top of appropriate APIs and evaluation standards (especially open source); keeping evaluations as model-agnostic as possible; facilitating model evaluation implementation in a privacy-preserving manner (e.g. without leaking additional information about a model to another party).

In addition, Signatories commit to reporting the level of uncertainty in their results and the limitations of the methods used. Ultimately, from the combination of these measures, the Signatories aim to adhere to the quality standards of scientific peer review in machine learning and the natural and social sciences, aiming to have key results for the most severe risks reviewed as thoroughly as expected from a submission to a major machine learning conference or journal.

Potential Key Performance Indicators

- KPI 10.3.1: Percentage of model evaluations used as evidence in a Model Report which show high internal and external validity (target: 75%).
- KPI 10.3.2: Percentage of model evaluations used as evidence for highest risk tier in a Model Report which have been reproduced (target: 50%).
- KPI 10.3.3: Percentage of model evaluations used as evidence in a Model Report which are fully portable (target for non-SMEs: 25%, target for SMEs: 0%).
- KPI 10.3.4: Percentage of model evaluations used as evidence in a Model Report where uncertainty is reported (target: 100%).

Measure 10.4. Model elicitation

Signatories commit to ensuring that evaluations of general-purpose AI models with systemic risk are being run with a state-of-the-art level of model elicitation, to appropriately elicit the capabilities, propensities and effects of a model under evaluation, minimise the risk of under-estimation, and match the realistic elicitation capabilities of potential misuse actors. This could include, for example, fine-tuning, prompt engineering, scaffolding (including tool use), using logits access, the ability to disable safeguards, or the use of base models or helpful-only models.

A state-of-the-art level of model elicitation is assumed when the combination of elicitation techniques used matches the best-known combination of techniques for the evaluated model, or comparable models, taking into account the best of both internally used, as well as externally documented and reproduceable techniques. Signatories commit to striving to use methods that minimise the risk of strategic model underperformance on an evaluation (i.e. sandbagging or strategic deception).

This requires providing evaluators with adequate compute budgets, to allow for long enough evaluation runs, parallel execution, re-runs, as well as appropriate staffing and engineering budgets to inspect results closely to, for example, identify software bugs or model refusals which might lead to artificially lowered estimates. For the most severe risks identified, Signatories commit to roughly matching the elicitation

efforts spent on their leading non-safety research projects. They may proportionally reduce these efforts for lower levels of risk.

In addition, Signatories commit to ensuring that elicitation methods with an increased risk profile (e.g. they make models better at harmful capabilities or increase harmful propensities) are matched with increased security measures (see Measure 12.2), to prevent proliferation risks or harmful actions by the elicited model.

Potential Key Performance Indicators

- KPI 10.4.1: Average percentage of engineering hours spent on model elicitation for model evaluations used as evidence for the highest risk tiers as compared to the largest internal non-safety project (target: 75%).
- KPI 10.4.2: Average percentage of engineering hours spent on model elicitation for model evaluations used as evidence for lower risk tiers as compared to the largest internal non-safety project (target: 33%).

Measure 10.5. Models as part of systems

Signatories commit to ensuring that model evaluations can assess the capabilities and limitations of a general-purpose AI model with systemic risk when integrated into an AI system. Signatories planning to use or integrate a general-purpose AI model with systemic risk into an AI system themselves will ensure that their model evaluations consider this future deployment situation, when relevant to the risk being assessed, and evaluate the model accordingly (for example by evaluating it using the same kind of scaffolding or tooling). If the deployment environment changes in a major way from what was originally used for assessment, they will take this into account and trigger a (partial or complete) model re-evaluation as defined in Measure 6.3.

Where Signatories do not use or integrate the general-purpose AI model in any AI system themselves (for example, through sole open-source release or sole B2B licensing), they will enforce their licensing terms to ensure that the licensee evaluates the model within an AI system representative of the AI system which they plan to use or deploy it in. This evaluation will be conducted to the same standard of rigour that the Signatory is applying to this model, particularly when assessing the highest risk tiers and considering the resources of the licensee. Signatories may provide licensees with the required scientific information and tooling needed for this model evaluation (see Measures 10.3 and 10.8).

Measure 10.6. Representative model evaluations & generalisation

Signatories commit to ensuring that model evaluations match the expected usage context of a model. For example, language-based evaluations of multilingual models may focus not only on English but will take into account major European languages (or other languages the model is claimed to support), as much as this is relevant to the risk being assessed (some risks are based on capabilities that might be best evaluated in English, i.e. those based on programming capabilities). Signatories commit to doing so not just for text but for all modalities supported by a model. For other modalities, representation might be measured in different categories than languages (e.g. image representations).

To facilitate this, Signatories are encouraged to make use of participatory methods involving expert or lay representatives of civil society, academia, and other relevant stakeholders, where this is appropriate for the risk being assessed. Where relevant stakeholders (e.g. those who are affected) are not available, Signatories commit to aiming to identify the best representatives who can represent their interests and consult them in the evidence gathering processes.

Measure 10.7. Exploratory work

Signatories commit to ensuring that exploratory work, appropriate to the size of their organisation and the risk being assessed, is done on their general-purpose AI models with systemic risk, such as exploratory research or open-ended red teaming (especially including the perspectives of civil society and other affected stakeholders). They commit to not restricting themselves only to evidence collection for known capabilities, but also to striving to assess risks from emerging capabilities, propensities, or effects through these methods.

Potential Key Performance Indicator

- KPI 10.7.1: Percentage of internal staff hours spent on exploratory safety research as part of all research hours (target for non-SMEs: 10%, target for SMEs and/or providers who exclusively release open-source models: 0%).

Measure 10.8. Sharing tools & best practices

Signatories commit to making state-of-the-art evidence collection best practices widely accessible to relevant actors in the AI ecosystem. They commit to also striving to make model evaluation tools available, where appropriate, especially for other model providers and third-party evaluators. Especially where Signatories have already developed significant in-house tooling and experience, they commit to aiming to share their expertise with SMEs and newer model providers.

Where appropriate, they may also share some model evaluation data (inputs and outputs), to further the safety work of other parties, but applying additional caution to model evaluation data that has high proliferation risks (i.e. the test set could become a training set for a misuse actor) or where it would threaten scientific rigour (i.e. leaking test data into training sets or fully releasing currently active held-out test sets). Where appropriate, these considerations will be documented in the Model Report. Signatories may also limit the sharing of information to protect public security, and commercially sensitive information, where these interests outweigh the safety benefit.

Signatories commit to ensuring that sharing the above will not come at the expense of their capacity to do safety research in the first place. They will consider hiring or assigning additional engineering and support staff to research teams to facilitate this work, while keeping the workload of existing safety researchers the same.

If this process is not (yet) facilitated by the AI Office or other public organisations designated by the AI Office (e.g. the international network of AI Safety Institutes), Signatories are encouraged to facilitate this sharing via industry organisations, but will ensure that there is no strict membership requirement or other restrictions which lead to the exclusion of potential competitors.

Potential Key Performance Indicators

- KPI 10.8.1: Number of relevant parties with whom evidence collection best practice guidance has been shared (target for non-SMEs: 10, target for SMEs: 0).
- KPI 10.8.2: Number of new model evaluation tools that have been shared with *at least 5* other providers or evaluators (target for non-SMEs: 3, target for SMEs and providers who substantially release open-source models: 0).

Measure 10.9. Qualified model evaluators and evaluation access

Signatories commit to ensuring that model evaluation teams possess the necessary qualifications for their work. This requires not only competence in machine learning but also access to the relevant domain expertise necessary to evaluate risks and other related aspects effectively.

Signatories commit to providing all model evaluators (internal and/or external) with the support needed to work to a rigorous scientific standard (as described in Measure 10.3), by giving them enough time, model access, engineering support, and compute budget to properly evaluate a general-purpose AI model with systemic risk. This also includes appropriate levels of model elicitation (see Measure 10.4) and access to models as part of systems where appropriate (see Measure 10.5). In addition, for external model evaluators, the conditions specified under Measure 16.1 apply.

To facilitate model elicitation, Signatories commit to working towards giving more model evaluators, internal and external, grey- and white-box access to their general-purpose AI models with systemic risk, considering a cost-benefit analysis with regards to model security (see Measure 12.2).

Measure 10.10 Safety margins

When Signatories use model evaluation results to inform the mapping of risk tiers to mitigations (see Measure 4.3), they commit to defining a considerable safety margin which is representative of potential under-elicitation, potential improvements after deployment, and general uncertainty in the evidence collection.

Measure 10.11 Forecast-ready model evaluations

Signatories are encouraged to make key model evaluations for the most severe risks more ‘forecast-ready’ over time, for example by running ‘scaling law experiments’ that intend to give predictions for when future models might be showing dangerous capabilities.

Potential Key Performance Indicators

- KPI 10.11.1: Percentage of evaluations used as evidence in Model Report, which allow for some level of forecasting (target: 10%).

Measure 10.12 Post-deployment monitoring

Signatories commit to conducting post-deployment monitoring for systemic risks at the highest risk tier and may conduct it for lower risk tiers as well. This includes, but is not limited to, the collection of relevant model-independent evidence defined in Measure 10.1.

Signatories commit to establishing contractual, technical, or other mechanisms to gather and include relevant post-deployment information in risk assessment and ensure the effectiveness of their mitigations. Relevant post-deployment information encompasses information Signatories can collect after placing the model on the market, which informs the identification and assessment of risks and mitigations. Such information includes but is not limited to information on model integration, model usage, and model impacts. These mechanisms may vary across different model integrations and usage and Signatories are encouraged to adapt their post-deployment monitoring to their model's distribution strategy and the type of customers and industries using the model.

All Signatories (including those who are providers of open-source models) are encouraged to consider methods such as: anonymous reporting channels; incident report forms and bug bounties; supporting community-driven evaluations and public leaderboards; monitoring evidence of model usage in the real world (for example identifying usage in software repositories and known malware, as well as monitoring public forums and social media for novel patterns of usage); supporting the scientific study of their models' emerging risks, capabilities, and effects in the market (using, for example, participatory methods); investing in new technical methods supporting the privacy-preserving monitoring and analysis of individual instances of AI agents built from their models (for example watermarks, fingerprinting, digital signatures, or decentralised, privacy-preserving monitoring techniques).

Signatories who are providers of closed-source models, deployed via an API, are encouraged to additionally monitor their general-purpose AI models with systemic risk via lightweight telemetry (logging) and data analysis methods, ensuring the preservation of user privacy and avoidance of transmitting sensitive data in accordance with EU laws and regulations. Where model providers themselves deploy AI systems, they will monitor these models as part of these systems, considering especially end-user feedback.

In addition, and especially where the above methods are not available and Signatories do not have access to enough relevant information via other means, Signatories commit to making use of, and enforcing, license agreements with downstream model deployers and users, which require those licensees to share – in a privacy preserving way – key evidence about systemic risks investigated by the Signatory. For end users who are natural persons signatories may allow those end users to opt in to sharing similar relevant information, ensuring the preservation of user privacy and avoidance of transmitting sensitive data in accordance with EU laws and regulations.

Potential Key Performance Indicators

- KPI 10.12.1: Percentage of risks at the highest risk tiers for which evidence is being collected using at least three forms of post-deployment monitoring (target: 75%).
- KPI 10.12.2: Percentage of risks at lower risk tiers for which evidence is being collected using at least one form of post-deployment monitoring (target: 50%).
- KPI 10.12.3 Percentage of licensees who report actionable evidence collected for the highest risk tiers (target: 50%).

Explanation of changes to Commitment 10 (Evidence collection): Most changes in this section were based on requests for clarifications on how certain key terms (such as "scientific rigour" or "state-of-the-art model evaluations") will be defined.

We also aimed to ensure that the options available to fulfil the Measures are proportional to the risks at hand (especially allowing for additional focus on the highest tiers of risks). Additionally, we took into consideration deployment contexts (for example, what methods are appropriate for open-source releases) and aimed to reduce bureaucratic burdens on SMEs while giving them as much clarity as possible. We also clarified that this Commitment can be applied to work done internally or externally (see also Commitment 16).

We also wanted to make explicitly clear that that post-deployment monitoring (which was moved here) was not prescribed as a fixed, inflexible and potentially inappropriate single technique but instead as a measure which could be fulfilled in ways that are most appropriate for the business model and deployment context around a model.

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

Commitment 11. Safety mitigations

In order to fulfil their obligations under Article 55(1), point (b) AI Act, Signatories commit to implementing safety mitigations that are proportional to the risks from the development and deployment of such models, and reducing risks to acceptable levels. Signatories commit to ensuring that safety mitigations are state-of-the-art and adhering to AI Office guidance where available.

In order to satisfy Commitment 11:

Measure 11.1. Safety mitigations to consider

Signatories are encouraged to consider implementing safety mitigations which include, but are not limited to: (a) filtering and cleaning training data, (b) monitoring and filtering the inputs and outputs of models, (c) changing the behaviour of a model, such as fine-tuning the model to refuse certain requests, (d) restricting the deployment of a model, such as restricting access of certain models to vetted users, (e) countermeasures or other safety tools made available to other actors to reduce systemic risk, and (f) high-assurance quantitative safety guarantees on the behaviour of a model.

Measure 11.2. State-of-the-art

State-of-the-art mitigations are those that best reflect the relevant findings of research, technology, and experience, and incorporate the best-known procedures for mitigating systemic risks. State-of-the-art mitigations are not always those that mitigate all risks to the greatest extent; instead, Signatories commit to striving to implement mitigations proportionate to the risks (cf. risk tiers in Measure 4.2). In this sense, state-of-the-art mitigations are those that best mitigate especially the highest-tier risks. Signatories are also encouraged to take into account any benefits lost by the implementation of mitigations.

Explanation of changes to Commitment 11: Previously, safety and security mitigations were covered by a single joint 'Measure 12'; in this draft, they are split into Commitments 11 and 12.

Some respondents felt there was a need for more detail on which safety mitigations are required. To give more detail, while also taking into account the fact that the field of technical risk mitigation is developing and has few established solutions that are known to work, we list a set of safety mitigations that Signatories are encouraged to consider.

We think that an outcome-based approach is likely most effective. Instead of requiring a fixed set of mitigations, which may be outdated (or worse, actively counterproductive) in the future, we think model providers will commit to implementing mitigations proportionate to the risk posed by a model.

Commitment 12. Security mitigations

In order to fulfil their obligations under Article 55(1), points (b) and (d) AI Act, Signatories commit to mitigating systemic risks from the possession of the unreleased weights, and associated assets such as unreleased algorithmic insights, of their general-purpose AI models with systemic risk. To do so, Signatories commit to implementing security mitigations designed to thwart attempts to obtain such weights and associated assets by highly motivated and well-resourced non-state actors, including insider threats, in line with the RAND SL3 benchmark. Signatories are encouraged to implement and advance research on more stringent security mitigations, in line with the RAND SL4 or above benchmarks, if they expect to face attempts to steal such weights or associated assets from state-level adversaries.

In order to satisfy Commitment 12:

Measure 12.1. General cybersecurity best practices

Signatories commit to adopting general cybersecurity best practices. This will include, but is not limited to:

(a) strong password policies and management,
(b) email filtering for suspicious links and phishing attempts,
(c) protection of wireless networks,
(d) policies for untrusted removable media,
(e) physical intrusion prevention, and
(f) regular software updates and patch management.

These best practices will be in line with cybersecurity technical standards such as ISO/IEC 27001, NIST 800-53, SOC 2, and a European harmonized standard on cybersecurity if available, and will also be in line with regulations such as the NIS2 Directive and the Cyber Resilience Act.

Measure 12.2. Security assurance

Signatories commit to implementing measures to assure and test their security readiness against the threat actors in Commitment 12. This will include, but is not limited to:

(a) frequent security reviews by an accredited third party,
(b) frequent active red-teaming,
(c) secure communication channels for third parties to report security issues,
(d) competitive bug bounty programs to encourage public participation in security testing,

(e) clear and public security whistleblower policies which prohibit retribution, in line with Commitment 18,

(f) installation of Endpoint Detection and Response (EDR) and Intrusion Detection System (IDS) tools on all company devices and network components, and

(g) the use of a security team to monitor for EDR alerts and perform incident handling, response and recovery for security breaches in a timely manner.

These measures will align with security assurance technical standards such as NIST 800-53 (CA-2(1), CA-8(2), RA-5(11), IR-4(14)), and NIST SP 800-115 5.2.

Measure 12.3. Protection of stored model weights and related assets

Signatories commit to implementing measures to protect stored model weights and associated assets such as unreleased algorithmic insights. This will include, but is not limited to:

(a) a registry of all devices and locations where model weights are stored,

(b) access control and monitoring of access on all devices storing model weights, with alerts on copying to non-controlled devices,

(c) storage on dedicated devices which host only data, code, and services treated with equivalent levels of sensitivity and security,

(d) ensuring model weights are always encrypted in storage and transport, with at least 256-bit security and with encryption keys stored securely on a Trusted Platform Module (TPM),

(e) ensuring model weights are only decrypted for legitimate use to non-persistent memory,

(f) implementing confidential computing once available and practical, using hardware-based, attested trusted execution environments to prevent unauthorised access to model weights while in use, and

(g) restricting physical access to data centres and other sensitive working environments to required personnel only, along with regular inspections of such sites for unauthorised personnel or devices.

These measures will be in line with technical standards such as NIST 800-53.

Measure 12.4. Interfaces and access control to model weights

Signatories commit to implementing measures to harden interfaces and access to model weights under storage or in use. This will include, but is not limited to:

(a) explicitly authorising only required software and personnel entities for both direct or indirect access to model weights, enforced through multi-factor authentication mechanisms, and audited on a regular basis;

(b) thorough review by a security team of any software interface accessing model weights for vulnerabilities or data leakage;

(c) hardening interfaces with access to the weights, to reduce the risk of data and weight exfiltration, using methods such as output rate limiting; and

(d) limiting the number of people who have non-hardened direct access to weights.

These measures will be in line with technical standards such as NIST SP 800-171, INCITS 359-2004, and NIST 800-53.

Potential Key Performance Indicator

- KPI 12.4.1: Percentage of defined interfaces and access points to unreleased model weights and associated assets that have implemented the specified security controls (target: 100%).

Measure 12.5. Insider threats

Signatories commit to implementing measures to screen for and protect against insider threats. This will include, but is not limited to:

(a) background checks on employees and contractors that have or might reasonably obtain access to unreleased model weights, associated assets, or systems that control the storage or use of unreleased model weights; and

(b) the provision of training on recognising and reporting insider threats.

These measures will align with technical standards such as NIST 800-53 (PM-12, PS-3).

Measure 12.6. Regime of applicability

Signatories commit to ensuring that the security measures in Measures 12.1—12.5 apply to all versions and copies of unreleased model weights and associated assets for general-purpose AI models with systemic risk throughout their lifecycle, from before training until any secure deletion of model weights or open releasing of them.

Potential Key Performance Indicator

- KPI 12.6.1: Percentage of copies and versions of unreleased model weights that have the security mitigations in Measures 12.1.–12.5. applied (target: 100%).

Measure 12.7. Limited disclosure

Signatories commit to ensuring that any publicly available copy of the Framework or the Model Report only discloses security mitigations in Measures 12.1—12.5 to a level of detail that does not undermine the effectiveness of the security mitigations and is consistent with publicly available details about such mitigations.

Potential Key Performance Indicator

- KPI 12.6.1: Percentage of copies and versions of unreleased model weights that have the security mitigations in Measures 12.1.–12.5. applied (target: 100%)

Explanation of changes to Commitment 12: Previously, safety and security mitigations were covered by a single joint ‘Measure 12’; in this draft, they are split into Commitments 11 and 12. In response to feedback about making mitigations more concrete, we specify a level of security mitigations that we believe to be appropriate for Signatories. We anticipate feedback on this level of security mitigation.

Commitment 13. Development and deployment decisions

In order to fulfil their obligations under Article 55(1), points (b) and (d) AI Act, Signatories commit to mitigating residual risks where safety and security mitigations are insufficient. To do so, Signatories

commit to establishing a process to decide whether and how to proceed with the development and deployment of a general-purpose AI model with systemic risk. This process will be described in the Framework and will include how decisions for this process are made and escalated through the organisation, including its management body (cf. Commitment 14).

In order to satisfy Commitment 13:

Measure 13.1. Conditions for not proceeding

Signatories commit to detailing in their Framework conditions under which further development and deployment of a general-purpose AI model with systemic risk will not proceed due to insufficient mitigations for (a) keeping risk below an unacceptable level, or (b) appropriately mitigating risk that is below an unacceptable level (cf. Measure 4.2). This will also describe a process that will be used for de-deploying a model (where not openly released) and falling back on another one, if needed based on any post-deployment monitoring and risk assessment. This commitment does not preclude the use of other conditions or processes used for decision-making for development and deployment.

Measure 13.2. Decision process for proceeding

Signatories commit to detailing in their Framework the decision process under which development and deployment of a general-purpose AI model with systemic risk can proceed, after a decision has been made to not proceed (cf. Measure 13.1). This will include the implementation of appropriate safety and security mitigations, followed by any necessary risk assessments to demonstrate that systemic risks are kept below an unacceptable level and appropriately mitigated below that level.

Measure 13.3. Staging deployment when proceeding

Signatories commit to detailing in their Framework whether and how the deployment of general-purpose AI models with systemic risk will be staged to keep systemic risks below an unacceptable level and appropriately mitigated below that. This may include practices such as (a) limiting API access to vetted users, (b) gradually expanding access based on risk assessments, (c) starting with a closed release before any open release, (d) using logging systems to track usage and safety concerns, (e) setting criteria for progressing through stages, including those based on risk tiers and user feedback, and (f) retaining the ability to restrict access.

Measure 13.4. Transparency into external input in decision-making

To provide transparency into decision-making around the management of unacceptable risks, Signatories commit to detailing in their Framework if and when development and deployment decisions will have input from external actors, including relevant government actors, particularly where unacceptable risks are involved. Fulfilling this commitment does not require such involvement or authorisation but does require transparency around its presence or absence, in alignment with existing international commitments such as the Frontier AI Safety Commitments (Outcome 3.VIII).

Explanation of changes to Commitment 13: Previously, Commitment 13 was numbered as ‘Measure 14’. We made two major changes in response to feedback: (a) providing more detail and examples about what

a description of a decision process should include, including a description of how deployment should be staged when deciding to proceed (the new Measure 13.3), and (b) making it clear that input into external decision-making is not required in this Commitment; we only require description if such input is required (Measure 13.4).

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

Providers of general-purpose AI models with systemic risk are required to mitigate possible systemic risks at Union level according to Article 55(1), point (b) AI Act. This Section sets out governance measures and processes to which Signatories commit in order to comply with this obligation. The Section also sets out measures that Signatories commit in order to fulfil the obligations in Articles 52(1), 53(1), point (a) (specifically Annex XI, Section 2 AI Act), 55(1), point (c), and 87 AI Act.

Commitment 14. Systemic risk responsibility allocation

In order to fulfil their obligations under Article 55(1) AI Act, Signatories commit to clearly allocating appropriate levels of responsibility and resources across their organisations, including within the management body – both in its management and supervisory functions – as well as in product, development, and other relevant operational teams, with a view to fulfilling their obligation to assess and proportionally mitigate systemic risks from their general-purpose AI models.

This includes allocating appropriate levels of responsibility and resources in line with their organisational complexity, governance structure, and the possible systemic risks posed by their general-purpose models with systemic risk. Commitment 14 includes allocating clear and appropriate responsibilities between those parts of the organisation focused on core business activities that may produce systemic risk; those overseeing, implementing, and supporting the organisation's systemic risk assessment and mitigation; and those providing independent assurance of the adequacy of risk assessment and mitigation.

Signatories are encouraged to promote a healthy risk culture, including setting the tone from the top, allowing effective communication and challenge, and appropriate incentives to discourage excessive risk-taking, including rewards for cautious behaviour and internal flagging of risks.

Potential Key Performance Indicators

- KPI 14.1: The management body in its management function
    - For SMEs: A specific individual in the executive team is allocated responsibility for the organisation's risk management efforts. This individual will have sufficient staffing, time, and support to carry out their duties.
    - For non-SMEs: In line with the Three Lines model (see e.g. Institute of Internal Auditors, 2024), the Signatory has appointed a Chief Risk Officer (or equivalent), who supports and oversees the Signatory's systemic risk management efforts. The Chief Risk Officer is supported by a sufficiently well-staffed risk function.
- KPI 14.2: The management body in its supervisory function
    - Specific individuals, such as in a risk or audit committee, have been given responsibility and adequate resources to oversee the organisation's management of systemic risk. This could be achieved through an external audit report provided to the relevant individuals.
    - The management body in its supervisory functions, or a suitable subset thereof, is given access to the information they need to oversee the organisation's systemic risk management, including, at minimum, all the information that the Signatory documents with

the purpose of providing it to the AI Office or national competent authorities upon request (see Commitments 1 and 20).

- For SMEs: A subset of the management body in its supervisory function is allocated responsibility for overseeing the organisation's risk management efforts and provided sufficient resources to carry out their task.
- For non-SMEs: In line with the Three Lines model (see e.g. Institute of Internal Auditors, 2024), a subset of the management body's supervisory function is allocated responsibility for and is sufficiently resourced for overseeing the organisation's systemic risk management efforts, for example by establishing an audit or risk committee. These individuals are supported by an internal audit function that reports directly to them, so as to provide them with independent evidence of the organisation's compliance and risk management efforts.

- KPI 14.3: The adherence and adequacy assessment (as per Commitment 15) concludes that appropriate levels of responsibility and resources have been allocated across the organisation, to assess and proportionally mitigate systemic risk.

Explanation of changes to Commitment 14: In response to stakeholder feedback, the language has been refined to accommodate different organisational structures, in acknowledgement of the fact that responsibility may be divided across various levels. The term 'management body,' in line with established usage, for example in the Digital Services Act Article 41, has been adopted. Measures have been consolidated into a single Commitment to make the framework adaptable to different governance models, including those without distinct board and executive functions. In addition, new KPIs and voluntary measures have been introduced to address stakeholder suggestions for robust oversight measures, such as internal audit functions and fostering a healthy risk culture.

Commitment 15. Framework adherence and adequacy assessment

To fulfil their obligations under Article 55(1) AI Act, Signatories commit to assessing their adherence to and adequacy of their Framework, the adoption and implementation of which they have committed to under Commitment 4.

Signatories commit to conducting an adherence assessment of general-purpose AI models with systemic risk, six months after placing them on the market. Signatories commit to conducting adequacy assessments of their Framework within four weeks of notifying the AI Office that it has or will meet the criteria in Article 51 AI Act, or every six months, whichever comes sooner. Signatories are encouraged to conduct such reviews more often, for example if there are significant changes in the risk landscape, or in response to serious incidents where their general-purpose AI model with systemic risk was deemed to play a significant role**.**

These assessments may be carried out by the Signatories' own staff (for example through an internal audit function, where applicable) or by external parties (as per Measure 16.1). Both the adherence and adequacy assessments will be reported to the management body in its supervisory function. The thoroughness of the reviews will be proportional to the possible systemic risk posed by the model in question.

The adequacy assessment will consider best practices, relevant research and state-of-the-art science, serious incident reports, and potential external expertise. It will cover the following questions, supported with sufficient evidence:

- Does the Framework address all components outlined in Commitment 4 with sufficient detail to evaluate its effectiveness? Is the Framework otherwise consistent with the Code?
- Does the Framework adequately account for uncertainty in future developments, as well as in the effectiveness in risk assessment and mitigation techniques?
- Is there strong reason to believe that the Framework will be adhered to over the next nine months? For example, does the Framework appear adequate for models planned to be placed on the market within this period?
- Are the policies, resources, and mitigations in the Framework still proportional to the complexity and scale of the systemic risks posed by the organisation's models which are planned to be placed on the market within the next nine months?

An adherence assessment should be conducted six months after putting a model on the market and cover the following questions, supported by sufficient evidence:

- Did the general-purpose AI model with systemic risk adhere to the relevant Framework?
- Did the relevant Model Reports adhere to the Framework? Did the Model Reports provide sufficient evidence to assess compliance with the Framework?
- Have post-market mitigations and assessments been implemented in accordance with the Model Report?
- Were any new risks documented in the Model Reports that were not previously identified in the Framework?

Potential Key Performance Indicators

- KPI 15.1: Adequacy and adherence assessments are conducted every six months.
- KPI 15.2: Adequacy and adherence assessments yield a positive result.
- KPI 15.3: Concerns identified during the adequacy and adherence assessments are addressed before the model is placed on the market or before the next scheduled adequacy and adherence assessment, whichever comes first.
- KPI 15.4: An external auditor provides an adherence or adequacy assessment with a positive result.
- KPI 15.5: The Framework is comparable in rigor, breadth, and depth to Frameworks used by other providers of general-purpose AI models with systemic risk with similar risk profiles.

Explanation of changes to Commitment 15: In response to stakeholder feedback, the provisions regarding adherence and adequacy assessments have been expanded to clarify their scope, methodology, and timing. These adjustments are not intended to establish new standards for the Framework or Model Report, but rather to confirm that existing frameworks and reports remain consistent with the Code and that evidence is collected to support their ongoing alignment. The purpose of the adequacy assessment is to help Signatories identify in what ways, if any, their current practices are insufficient, giving them time to adjust during the development of the model.

Commitment 16. External risk assessment

To fulfil their obligations under Article 55(1) AI Act, Signatories commit to enabling sufficient external expert risk and mitigation assessments of general-purpose AI models with systemic risk throughout their lifecycle, as appropriate, with a view to fulfilling their obligation to assess and mitigate possible systemic risks from their general-purpose AI model according to Commitments 6—12.

In order to satisfy Commitment 16:

Measure 16.1 Before market placement

Signatories commit to having qualified external assessors conduct relevant parts of risk assessment pursuant to Commitments 6—10 before market placement (as per Measures 6.2 and 6.3) for their general-purpose AI model with systemic risk. This commitment is limited to a subset of the systemic risks identified according to Commitment 7, for which the following applies:

- The provider cannot provide sufficient evidence that the model does not pose additional risk beyond that of general-purpose AI models with systemic risk already on the EU market, for example due to novel model capabilities or differences in implemented mitigations (similarity to a model already on the EU market, as defined in Commitment 10, would constitute such evidence);

  or

- the Signatory has insufficient relevant internal expertise or information to effectively conduct the risk assessment (such as those assessments requiring sensitive national security-relevant information).

Signatories commit to including an explanation of whether the above criteria have been met in their Model Report. If the criteria above are met, but Signatories fail to identify qualified external assessors, as per the criteria below after a good faith effort (including early search efforts and, if a qualified assessor is found, early notification of external evaluators as per Measure 6.1), they commit to relying on internal assessments and providing a justification of the lack of external assessment in their Model Report.

Signatories are encouraged to involve qualified external assessors in holistically assessing their Framework and Model Reports, in particular, assessing their alignment with best practices with regards to their risk assessment and mitigation efforts.

Signatories commit to taking information gathered via external assessments, where available, into account when deciding to place a model on the market and including such assessments in their Model Report. If Signatories disagree with any recommendations or conclusions reached by external assessors, they commit to specifying their disagreement in the Model Report.

Assessments may be conducted by the AI Office or other qualified external assessors, including government bodies, that have been recognised by the AI Office or meet the criteria below:

- have significant domain expertise for the risk domain(s) or risk model(s) they are evaluating for,

- are technically skilled and experienced in conducting model evaluations, if relevant,
- have internal and external information security protocols in place, sufficient for the level and type of access granted,
- have sufficient levels of independence from the provider, including safeguards such as contractual provisions ensuring impartiality in their results, including appropriate conflict of interest policies.

Where Signatories do involve external assessors, they commit to including a justification of their evaluator choice in their Model Report, based on the qualification criteria outlined above. A justification is not required where the external assessors have been recognised by the AI Office, where applicable.

Where external assessors conduct model evaluations (as per Commitment 10), Signatories commit to ensuring such testing involves, among other things, sufficient model access, information about the model, time, resources (as per Measure 10.9), and non-retention of experiments, as well as the ability to independently report results and guarantees of non-retaliation, including restrictions of model access.

Signatories are encouraged to use the evidence generated through such external assessments to support them in their responsibility to provide evidence for Commitments 6—10.

When do Signatories commit to external assessment of a specific systemic risk *pre-deployment*?

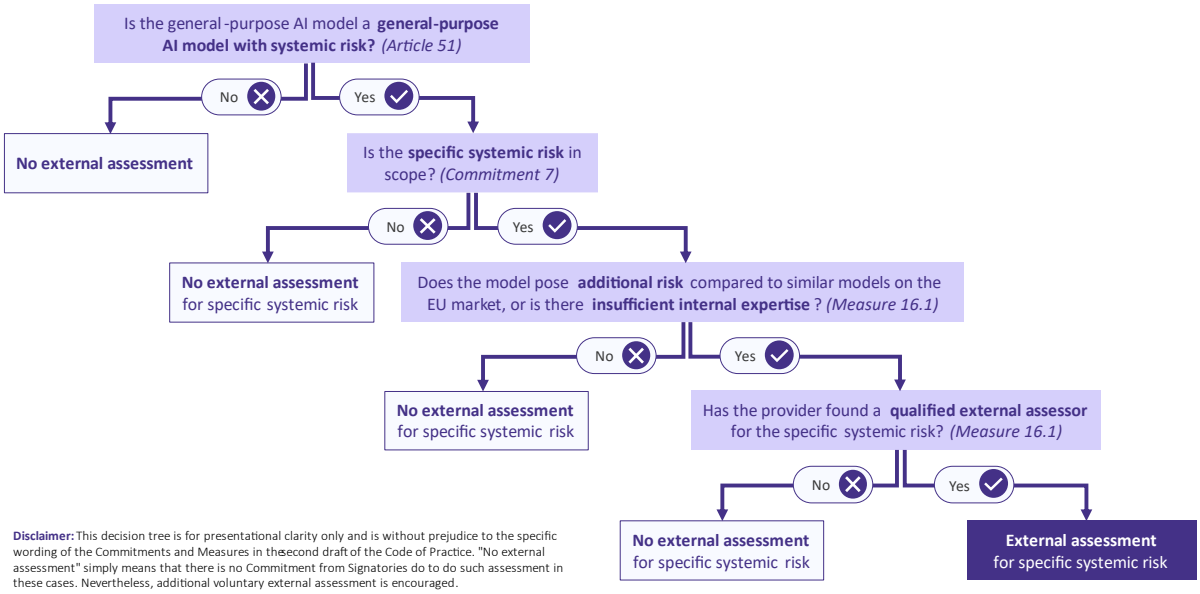

Figure 1: Decision tree outlining the conditions under which Signatories commit to conducting an external assessment of a specific systemic risk pre-deployment

Explanation of the changes to Measure 16.1: This revision is a direct response to feedback received from multiple stakeholder groups, with some calling for stricter external assessment requirements than what was proposed in the first draft, and others arguing that the Commitment should be entirely voluntary. All stakeholders requested significantly more detail on what 'adequacy' means regarding external assessors; the updated text now clarifies who qualifies as a 'qualified' external assessor and

points to continually updated guidance, including a list of recognised external assessors from the AI Office.

Providers, in particular, voiced concerns about strict mandates for external assessments given that the ecosystem remains nascent, expressed fears that Measure 16.1 could lead to a de facto pre-approval regime, and concerns that such external assessment might be unnecessary and onerous. At the same time, other stakeholders continue to strongly emphasise the importance of sufficient, mandatory external testing before deployment. To address these concerns, the revision clarifies in which specific cases external assessments is appropriate, and in which cases such an external assessment is not necessary. This is in line with our interpretation of Recital 114 AI Act. It further retains flexibility by allowing providers to justify, under specific circumstances, why they cannot conduct external assessments, even where it would be appropriate. Further, by ensuring that external assessment results are submitted in the Model Report which is only due at the time of deployment (but not before) and that providers are free to disagree with recommendations made by external assessors (where they are made), we have tried to clarify that providers need not follow them. Finally, the text opens the possibility of mutual recognition of external assessment results across international government bodies and external assessors, so long as they fulfil the criteria for qualified assessors and adhere to the risk assessment standards outlined in Commitments 6—10; in particular, it does *not* state EU residency as a requirement to be considered a qualified assessor.

Measure 16.2. After market placement

Signatories commit to ensuring that general-purpose AI models with systemic risk are subject to sufficient exploratory assessment by external assessors to identify, assess, and mitigate systemic risks after market placement. This will take into account guidance from the AI Office on conducting exploratory post-deployment assessment by external assessors, where available. For Signatories who are providers of closed-source models, this commitment includes creating a so-called safe harbour regime for external assessors, which includes facilitating secure and non-restrictive access to deployed models for independent model evaluation, subject to adherence to established rules of engagement and safeguards to prevent misuse. Such access protocols must balance the need for robust model evaluation with the requirement to maintain the integrity and confidentiality of the models and evaluation environments. Releasing a model under a free and open-source licence is another way to provide such access. Where models are released in an open-weight, but not open-source manner, Signatories commit to facilitating research conducted on their model.

Measure 16.2 includes supporting the independence of experts by refraining from imposing restrictions on the assessors or their publication of findings, provided these findings align with a responsible disclosure framework. Where external assessors submit reports outlining identified risks, testing methodologies, and proposed mitigations to Signatories, Signatories commit to including them in their Model Reports. Signatories are further encouraged to use such evidence to fulfil Commitments 6.4 and 10.2.

Signatories are encouraged to actively incentivise external assessment after market placement, such as via bug bounty programs.

Potential Key Performance Indicators

- KPI 16.2.1: The Signatory has documented and implemented/published secure and non-restrictive access protocols for external assessors, including rules of engagement, principles for responsible disclosure, and reporting requirements.
- KPI 16.2.2: The Signatory has established bug bounty programs with clear success criteria, financial or compute-based rewards, as well as appropriate guarantees of non-retaliation.
- KPI 16.2.3: Independent external assessors, under Measure 16.2, receive access to deployed models within 30 days of a formal request, barring exceptional circumstances.
- KPI 16.2.4: The Signatory has documented a policy that permits external assessors to publish findings after the 60-day notification period, provided the findings adhere to the responsible disclosure framework.
- KPI 16.2.5: The Signatory has included all independent test results in their Model Report.
- KPI 16.2.6: The Signatory has adopted a safe harbour policy to protect external assessors from legal or financial consequences when acting in good faith and in compliance with the rules of engagement.

Explanation of changes to Measure 16.2: Few respondents objected to this Measure in the first draft. As such, in this draft we aimed to clarify the Measure further, by outlining what a safe harbour might entail.

Commitment 17. Serious incident reporting

LEGAL TEXT

Article 55(1)(c) AI Act: "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;"

In order to fulfil their obligations under Article 55(1), point (c) AI Act, Signatories commit to setting up processes with adequate resourcing.

Signatories understand what constitutes undue delay relative to the seriousness of the incident, the extent to which the incident is ongoing, and whether the AI Office or other external parties can respond to it. For ongoing incidents, Signatories commit to particularly avoiding delay in reporting.

A serious incident is defined in accordance with Article 3(49) AI Act, in the absence of other AI Office guidance. However, Signatories commit to focusing their identification efforts on serious incidents that are more feasible to identify, such as: the death of a person, or serious harm to a person's health (Article 3(49), point (a) AI Act), a serious and irreversible disruption of the management or operation of critical infrastructure (Article 3(49), point (b) AI Act), and serious harm to property or the environment (Article 3(49), point (d) AI Act).

LEGAL TEXT

Article 3(49) AI Act: "'serious incident' means an incident or malfunctioning of an AI system that directly or indirectly leads to any of the following:
(a) the death of a person, or serious harm to a person's health;
(b) a serious and irreversible disruption of the management or operation of critical infrastructure.
(c) the infringement of obligations under Union law intended to protect fundamental rights;
(d) serious harm to property or the environment;"

This commitment includes allocating sufficient resources to investigate any suspicion of their model's involvement, direct or indirect, in a serious incident. It further includes that if the suspicion is confirmed, Signatories commit to reporting the incident without undue delay to the AI Office. A Signatory reporting an incident to the AI Office does not constitute an admission of wrongdoing. Similarly, Signatories commit to not treating serious incident reports as finalised documents; they commit to retaining and continuing to add evidence and update the reports where appropriate.

Signatories commit to using appropriate methods to identify serious incidents, appropriate given their business models and means of deployment. They may for example:

- Facilitate third-party serious incident reporting, either directly to the AI Office and/or national competent authorities, or indirectly via the Signatory.
- Facilitate serious incident reporting from downstream providers and users, such as API customers, either directly to the AI Office and/or national competent authorities, or indirectly via the Signatory.
- Assess user logs for concerning queries and generations, where applicable.
- Identify content produced by their model using watermarks or other techniques for tracing output back to a model.
- Review other sources for evidence on incidents, such as police and media reports, posts on social media, dark web forums, and research conducted by external parties.

Signatories commit to triaging their serious incident identification efforts, spending more resources on incidents that are of higher severity and where there is stronger evidence of the model being indirectly or directly involved. Signatories commit to retaining documentation produced in the process of identifying, reporting, and responding to serious incidents for at least 12 months.

Signatories commit to following guidance and templates on serious incident reporting provided by the AI Office where available. In their absence, Signatories commit to reporting the following elements:

- Start and end date of the incident, or an approximation thereof.
- Nature of the incident, including the resulting harm.
- A description of the incident, including the chain of events that lead to it.
- Root cause analysis, including, as far as possible, a description of the model's outputs and the factors that led to their generation. This includes identifying the inputs used and any potential failures or circumvention of safeguards that contributed to producing these outputs.
- The model that was potentially involved in the incident, be it directly or indirectly.

- A description of evidence available regarding the model's direct or indirect involvement with the serious incident.
- What, if anything, the Signatory intends to do or has done in response.
- What, if anything, the Signatory recommends the AI Office, or national competent authorities do in response.

Where the Signatory does not yet have certain relevant information, they may note that in their serious incident report. Where an incident has already been reported via another framework, such as NIS2, Signatories may send that report instead.

Signatories are encouraged to also report near-misses, where a serious incident was close to occurring, in line with guidance from the AI Office, where available.

Explanation of changes to Commitment 17: Respondents were positive about the Commitment, but requested clarity on what constitutes a serious incident, how they will be identified, and how they will be reported. We've tried to add clarity to these points.

Further, there was a discrepancy between Commitment 17 and Measure 17.1 in the first draft, where the latter discussed "near misses". Reporting of near misses has now been added as an "encouraged" action.

Some respondents thought it would be best to remove Measure 17.2 in the first draft "Serious Incident Response Readiness," seeing as it would fit better into other parts of the Code. In response, we've moved that Measure into Measure 4.5 in this draft.

Commitment 18. Whistleblowing protections

LEGAL TEXT

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

In order to fulfil their obligations under Article 87 AI Act, Signatories with more than 50 employees commit to implementing whistleblowing channels and afford appropriate whistleblowing protections to covered persons and activities, as per Directive (EU) 2019/1937.

Signatories recognise their obligations regarding whistleblowing protections following from Directive (EU) 2019/1937. This recognition is without prejudice to the full application of these obligations and how they have been transposed into national law. The following provisions are intended to highlight the most significant obligations and do not constitute an exhaustive list of all responsibilities under Directive (EU) 2019/1937:

- establishing secure and confidential internal reporting channels that allow for both written and oral reporting.

- designating an impartial person or department to handle reports, acknowledge receipt within seven days, and provide feedback within three months.
- maintaining proper records while ensuring confidentiality and respecting data protection requirements including but not limited to the specification under GDPR.
- providing clear information about both internal and external reporting procedures.
- protecting whistleblowers from all forms of retaliation, including dismissal, demotion, or harassment.

Should there be a designated mailbox for whistleblowing set up by the AI Office, Signatories commit to informing covered persons of its existence.

Signatories with less than 50 employees are encouraged to implement the whistleblower protections outlined above, in particular, at least annually informing employees of an AI Office mailbox, if it is operational, and implementing measures and policies to protect whistleblowers from all forms of retaliation, including dismissal, demotion, or harassment.

Potential Key Performance Indicators

- KPI 18.1: Documented evidence of a whistleblower policy that ensures employees can report wrongdoing freely, anonymously, and without fear of retaliation. This policy will include:
  - Scope of protected disclosures.
  - Internal and external reporting channels.
  - Anti-retaliation measures and assurances.
  - Investigation process for reported concerns.
  - Jurisdictional affordances and limitations of protections.
- KPI 18.2: Description of the communication strategy for informing employees and stakeholders about whistleblower protections and reporting channels.
- KPI 18.3: Evidence of regular awareness initiatives (e.g. training, internal communications, campaigns) conducted at least annually.
- KPI 18.4: Provision of anonymous internal reporting mechanisms (e.g. compliance hotlines, secure messaging platforms) specifically for concerns related to general-purpose AI models with systemic risk.
- KPI 18.5: Documentation of AI-related whistleblower reports received, categorised by concern type, while maintaining confidentiality.
- KPI 18.6: Documentation of initial response times and the time taken to act upon the communicated concern. The initial response and acknowledgement of the whistleblowing report will happen without undue delay, while actions taken in response to the whistleblower report will be proportionate to the severity of the concerns reported.

Explanation of the changes to Commitment 18: A variety of stakeholders gave the feedback that the most important obligations under Directive (EU) 2019/1937 should be reiterated here in the Code. In addition, language was added that encourages small companies that don't fall under the Directive to implement an adjusted minimum version of whistleblowing protections.

Commitment 19. Notifications

In order to fulfil their obligations under Article 52(1) and Article 55(1) AI Act, Signatories commit to notifying the AI Office of relevant information regarding their models meeting the threshold for general-purpose AI models to classify as general-purpose AI models with systemic risk, their Framework, their Model Report, 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 according to Article 78 AI Act.

In order to satisfy Commitment 19:

Measure 19.1 General-purpose AI model with systemic risk notification

Signatories commit to, before starting a training run, estimating the amount of computational power they intend to use. If that computational power exceeds the threshold laid down in Article 51(2) AI Act, the Signatories commit to notifying the AI Office within two weeks. Signatories may choose not to notify the AI Office of a model that meets the classification criteria if they have strong reason to believe they will not put it on the EU market.

Computational power estimates will be done in accordance with best practice or guidance from the AI Office where available.

When notifying the AI Office that they will place a model on the market which will be classified as a general-purpose model with systemic risk, Signatories commit to also specifying whether they consider their current Framework adequate for the relevant model, if adhered to.

Explanation of the changes to Measure 19.1: Measure 19.1 primarily clarifies when a Signatory is expected to know that the condition in Article 51(1), point (a) AI Act, will be met. There was some concern that this would impose an additional obligation on providers to estimate the compute they will use to train models. However, our understanding is that these training runs involve substantial costs and resources, so it is unlikely that companies undertake them without a reasonable estimate of what they will require. The changes reflect this understanding. Furthermore, the changes also acknowledge that some models may be trained solely for internal purposes or as preliminary experiments, with no intention of being placed on the EU market.

Providers also expressed concern about the lack of clear standards for estimating compute usage during model training. We therefore clarified that such estimates should be done according to standards where they do exist.

The measure may need to be updated to account for changes that the Office might make to the classification criteria for general-purpose AI models with systemic risk as per 51(3).

Measure 19.2. Framework update notification

Signatories commit to ensuring the AI Office has access to the latest unredacted version of their Framework, in a timely manner and not later than within five business days of a confirmed update. The process for confirming updates will be outlined in the Framework. Such access could be provided via a publicly accessible link or via a sufficiently secure channel specified by the AI Office.

Measure 19.3 Framework adequacy assessment notification

Signatories commit to ensuring the AI Office has access to the latest unredacted version of their Framework adequacy assessment, in a timely manner and not later than within five business days of a confirmed assessment.

Measure 19.4. Safety and Security Model Report notification

Signatories commit to sending a Safety and Security Model Report to the AI Office by the time they place a general-purpose AI model with systemic risk on the market, through sufficiently secure channels specified by the AI Office. Signatories are encouraged to share their Safety and Security Model Report two weeks before the placement of a general-purpose AI model with systemic risk on the market.

Explanation of changes to Measure 19.4: Providers were concerned that submitting a Safety and Security Model Report to the AI Office might suggest requiring approval before deployment. To address this, we clarified that the Model Report is due at the time of market placement, not before.

Commitment 20. Documentation

In order to fulfil their obligations under Article 53(1), point (a), and Article 55(1) AI Act, Signatories commit to documenting information relevant to their adherence to the Code, throughout the model lifecycle, for the purposes of providing this information, upon request, to the AI Office and national competent authorities. Where available, the Signatories commit to making use of reporting templates or APIs provided by the AI Office, and otherwise work with standardisation, government, or industry organisations to develop and follow such templates to make it as easy as possible to compare results from different providers, and to allow for automated analysis of included data, as appropriate. Signatories commit to promptly responding to requests for documentation by the AI Office and offering clarifications where requested, be it via further documentation or interviews.

Measure 20.1. Documentation regarding classification based on Article 51 AI Act

Signatories commit to documenting information relevant to the classification of general-purpose AI models with systemic risk based on Article 51 AI Act.

More specifically, this entails:

- The documentation requirements outlined in Commitment 1.
- Results from risk assessments, described in Commitments 6—10, including model evaluation results, and an assessment of the model's generality, autonomy, and ability to use and access tools and scaffolding (see Annex XI, Section 2, point 1. and Annex XIII AI Act). This may be covered in the Model Report.
- The reach of the model, in particular the number of registered business users established in the Union (Annex XIII, point (f) AI Act).

Measure 20.2. Documentation regarding adherence to the Code and the AI Act

Signatories commit to documenting information relevant to their adherence to the Code. This includes:

- Their Framework.

- Their Model Reports.
- Adequacy and adherence assessments described in Commitment 15.
- Where applicable, a detailed description of the system architecture explaining how software components build or feed into each other and integrate into the overall processing (Annex XI, Section 2, point 3. AI Act), including input/output filters, fine-tuning, system prompts, feature inhibition/activation, query routing, hidden chains of thought, and tool use.
- Evidence of a whistleblower policy and of informing employees and stakeholders about whistleblower protections and reporting channels.

Explanation of changes to Commitment 20: This Commitment was clarified to ensure consistency and avoid duplication with the obligations that providers of general-purpose AI model already face under Article 53 AI Act, further detailed in Commitment 1. Given the complexity of the documentation requirements, there's a chance we've double counted in places or missed something; if this is the case, please let us know in the EU survey.

Commitment 21. Public transparency

Signatories commit to offering appropriate public transparency with the aim of aiding the wider ecosystem, including the public and external researchers, to better understand and mitigate systemic risks stemming from general-purpose AI models, especially in light of the nascency of the science of assessing and mitigating systemic risks. At a minimum, Signatories commit to publishing their Framework and their Model Reports or similar documents (such as model cards or system cards). Information may be redacted where its inclusion would substantially increase systemic risk, such as the disclosure of information about safety mitigations that would undermine their effectiveness or divulge sensitive commercial information to a degree disproportionate to the societal benefit.

Potential Key Performance Indicators

- KPI 21.1: Signatories publish updated or new versions of their Frameworks within 15 working days of sending it to the AI Office, potentially redacting information where its inclusion would substantially increase systemic risk or divulge sensitive commercial information to a degree disproportionate to the societal benefit.
- KPI 21.2: Signatories publish Model Reports within five days after the relevant model is released on the EU market, redacting information where its inclusion would substantially increase systemic risk or divulge sensitive commercial information to a degree disproportionate to the societal benefit.
- KPI 21.3: Signatories publish documents summarising the contents of their Model Report within five days of the relevant model being released on the EU market. For example, they publish model or system cards describing the results from systemic risk assessments, as well as mitigations put in place.

Explanation of changes to Commitment 21: This Commitment was revised following provider feedback that publishing Model Reports was overly burdensome and goes beyond the requirements of the AI Act. To address concerns, we eased disclosure requirements, allowing the redaction of sensitive commercial information where societal benefits are minimal. We also clarified that model or system cards could be substitutes for a summarised or redacted version of Model Reports, something which is typically published alongside general-purpose AI models with systemic risk.