AI & Partners

Amsterdam - London - Singapore

EU AI Act

Methodology

Classification of General-Purpose AI Models as General-Purpose AI Models with Systemic Risks

December 2024

AI & Partners

Sean Musch, AI & Partners

Michael Borrelli, AI & Partners

Charles Kerrigan, CMS UK

AI & Partners defends and extends the digital rights of users at risk around the world. By combining direct technical support, comprehensive policy engagement, global advocacy, grassroots professional services, regulatory interventions, and participating in industry groups such as AI Commons, we fight for fundamental rights in the artificial intelligence age.

This report was prepared by **Sean Musch**, AI & Partners, **Michael Charles Borrelli**, AI & Partners, and **Charles Kerrigan**, CMS UK. For more information visit https://www.ai-and-partners.com/.

Contact: AI & Partners | contact@ai-and-partners.com.

Contents

Executive Summary

The "Methodology for Classifying General-Purpose AI Models with Systemic Risks" under the European Union Artificial Intelligence Act (EU AI Act) offers a comprehensive framework to address the challenges posed by rapidly advancing general-purpose AI (GPAI) technologies. These models demonstrate wide-ranging capabilities across domains and hold transformative potential for innovation, but their scale, functionality, and societal influence also pose systemic risks. The document establishes a structured methodology to identify, classify, and mitigate these risks while ensuring AI models are deployed ethically, responsibly, and sustainably.

Central to the framework is the classification of GPAI models as systemic risk-bearing entities. A key threshold involves evaluating the computation used in training these models, quantified as exceeding 102510^{25}1025 floating-point operations (FLOPs). This computation measure reflects the advanced capabilities and complexity of such models, which necessitate careful scrutiny. Additionally, models with "high-impact capabilities," such as autonomy, scalability, and potential to influence societal or economic systems, are flagged for further assessment. Such capabilities indicate a model's propensity to disrupt critical sectors like healthcare, finance, and public safety, either through misuse or unintentional consequences.

To ensure a robust assessment, the methodology integrates multiple evaluation factors, including the quality of training data, model size and parameters, adaptability to novel tasks, and market reach. A model accessible to over 10,000 business users in the EU market is presumed to have significant systemic impact. Advanced technical protocols, such as adversarial testing and continuous monitoring, are employed to identify vulnerabilities and maintain compliance with regulatory standards. Documentation, transparency, and real-time risk reporting are emphasized to uphold accountability and facilitate timely interventions.

The methodology underscores continuous risk management. AI providers must notify the European Commission if their models meet systemic risk criteria, enabling proactive regulatory oversight. They may also submit arguments to challenge the classification by presenting evidence of mitigating factors, such as restricted usage or built-in safeguards. If systemic risks are confirmed, designated GPAI models must adhere to additional obligations, including enhanced monitoring, security measures, and periodic reassessments.

The dynamic reassessment mechanism reflects the evolving nature of AI. Providers can request a re-evaluation of their model's systemic risk designation if new evidence or advancements mitigate earlier risks. This adaptive approach ensures fairness and encourages innovation while safeguarding societal interests.

The document also highlights specific considerations for open-source GPAI models. While these models may benefit from certain exemptions, those classified as systemic risks are subject to the same stringent requirements as proprietary models. The balance between fostering open innovation and mitigating risks remains a critical focus.

By outlining these measures, the methodology aligns with the EU AI Act's overarching goals of fostering safe, ethical, and trustworthy AI innovation. It ensures that the benefits of AI technologies are maximized while mitigating risks to public health, security, fundamental rights, and societal well-being. Through this approach, the framework sets a global standard for AI governance, supporting technological progress while maintaining accountability and public trust.

Introduction

The rapid advancement of general-purpose AI (GPAI) systems has introduced technologies capable of performing a wide range of tasks across diverse domains. These models, while offering immense potential for innovation and efficiency, also pose significant systemic risks. These risks stem from the scale and impact of GPAI capabilities, particularly when such models achieve levels of performance comparable to or exceeding those of the most advanced AI systems. Recognizing and managing these risks is crucial to ensuring the safe, ethical, and sustainable deployment of AI technologies.

This document outlines a robust methodology for the classification of general-purpose AI models as possessing systemic risk. The framework is built around clear **classification criteria** and evaluation processes to identify and address risks effectively. Key among these criteria is the presence of **high-impact capabilities**, which signal a model's potential to influence or disrupt critical societal, economic, or technological systems. Additionally, the framework leverages a quantitative threshold: models requiring cumulative computation exceeding 102510^{25}1025 floating point operations during training are presumed to present systemic risks, as this threshold reflects significant computational and functional capabilities.

Beyond classification, the methodology emphasizes a set of **evaluation factors** to provide a comprehensive assessment of GPAI models. These factors include the quality and size of the training data, the scale and complexity of the model's parameters, and the computational resources used during development. The ability of a model to scale across modalities, such as text-to-image or text-to-text, and to autonomously adapt to new tasks, is also critical. Furthermore, the methodology evaluates a model's **market impact**, defined by its availability to at least 10,000 registered business users within the EU, to gauge its systemic reach and influence.

The methodology incorporates **technical tools and continuous monitoring** to support rigorous evaluation and mitigation efforts. State-of-the-art protocols, including adversarial testing, are deployed to uncover vulnerabilities and potential risks. Continuous risk assessment ensures that emerging issues are promptly addressed, and serious incidents are documented and reported to regulatory authorities.

Finally, the framework provides a **reassessment process** to adapt to evolving AI capabilities. Providers can request reassessment if new evidence arises, and the Commission retains authority to reassess models based on updated criteria.

By providing this structured approach, the methodology fosters transparency, accountability, and the responsible management of systemic risks, ensuring that GPAI models contribute positively to society while minimizing potential harms.

Objectives and Scope

The objective of this methodology is to establish a clear and comprehensive framework for identifying GPAI models that present systemic risks. As the capabilities of GPAI systems expand, so do the potential consequences of their misuse or malfunction. These risks can range from unintended societal impacts to significant disruptions in economic or technological systems. The classification process aims to identify and mitigate these risks, ensuring the responsible development, deployment, and monitoring of GPAI systems in alignment with ethical and regulatory standards.

Objectives

Identify High-Impact Capabilities:

The methodology focuses on assessing whether a GPAI model demonstrates high-impact capabilities comparable to or exceeding those of the most advanced AI systems. Such capabilities indicate a heightened potential for systemic risks and necessitate thorough evaluation and regulation.

Establish Computation Thresholds:

By defining a computation threshold of

floating point operations, the methodology provides a quantifiable measure to determine whether a model's training indicates systemic risk potential. This threshold serves as a proxy for the model's complexity and capability.

Comprehensive Evaluation:

The framework ensures a detailed evaluation of factors influencing a model's risk profile, including training data quality, parameter size, computational resources, and adaptability across tasks and modalities.

Promote Continuous Monitoring and Mitigation:

The methodology underscores the importance of ongoing risk assessment, incident reporting, and mitigation strategies to adapt to evolving risks throughout the model's lifecycle.

Provide for Reassessment:

Recognizing the dynamic nature of AI development, the methodology includes provisions for reassessment, ensuring that decisions remain relevant as new evidence or circumstances arise.

Scope

The scope of this methodology encompasses all general-purpose AI models that are developed, deployed, or made available within the European Union (EU). It is designed to apply across various stages of a model's lifecycle, from training and deployment to market impact and post-market monitoring.

Classification Scope:

The methodology applies to GPAI models that meet specific classification criteria, particularly those demonstrating high-impact capabilities or surpassing the

floating point operation threshold during training. It also considers the model's influence on the internal market, with a significant impact presumed for models accessible to at least 10,000 registered EU business users.

Evaluation Factors:

The framework's evaluation process covers a range of technical and functional aspects, including:

- Data and Parameters: Training dataset quality and model size.
- Computation and Modality: Resource intensity and flexibility across input/output modalities, such as text-to-text or text-to-image.
- Autonomy and Scalability: The model's ability to independently adapt to new tasks without retraining.
- Market Impact: The model's presence and influence within the EU market.

Tools and Monitoring:

The methodology requires the use of standardized protocols, such as adversarial testing, to identify vulnerabilities and mitigate risks. Continuous monitoring mechanisms ensure prompt detection of emerging issues and compliance with regulatory obligations.

Reassessment Process:

Providers can request reassessment if new, objective information arises, ensuring adaptability to changing circumstances. Additionally, the Commission retains the authority to reassess models based on updated criteria, such as changes in data quality or user impact.

By focusing on these objectives and encompassing a wide scope, this methodology provides a structured approach to identify and address systemic risks in GPAI models, contributing to a safer and more responsible AI ecosystem.

Definitions

AI Act Term	AI Act Definition
AI System	A machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions, that influence physical or virtual environments.
Authorised Representative	Any natural or legal person located or established in the EU who has received and accepted a mandate from a Provider to carry out its obligations on its behalf.
Deployer	A natural or legal person, public authority, agency, or other body using an AI system under its authority.
Distributor	Any natural or legal person in the supply chain, not being the Provider or Importer, who makes an AI System available in the EU market.
General-Purpose AI Model ("GPAI")	Means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market;
High-impact capabilities	Means capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models.
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
General-purpose AI system	Means an AI system which is based on a general-purpose AI model and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems
Floating-point operation	Means any mathematical operation or assignment involving floating-point numbers, which are a subset of the real numbers typically represented on computers by an integer of fixed precision scaled by an integer exponent of a fixed base.
Importer	Any natural or legal person within the EU that places on the market or puts into service an AI system that bears the name or trademark of a natural or legal person established outside the EU.

Operator	A general term referring to all the terms above (Provider, Deployer, Authorised Representative, Importer, Distributor, or Product Manufacturer).
Product Manufacturer	A manufacturer of an AI System that is put on the market or a manufacturer that puts into service an AI System together with its product and under its own name or trademark.
Provider	A natural or legal person, public authority, agency, or other body that is or has developed an AI system to place on the market, or to put into service under its own name or trademark.
Trustworthy AI	Defined through a set of principles aimed at ensuring that AI systems are developed and used in a manner that is ethical, respects fundamental rights, and is aligned with societal values. These principles, as outlined in the references provided, include: • Human Agency and Oversight: AI systems should serve people, respect human dignity, personal autonomy, and can be overseen and controlled by humans. • Technical Robustness and Safety: AI systems should be resilient and secure, minimizing unintended harm and ensuring reliability. • Privacy and Data Governance: Development and use of AI should comply with privacy and data protection rules, ensuring data quality and integrity. • Transparency: AI systems should be transparent, providing traceability and explainability, making users aware of AI interaction, and informing deployers and affected persons about their rights. • Diversity, Non-discrimination, and Fairness: AI development and use should promote equal access, gender equality, cultural diversity, and prevent discriminatory impacts. • Societal and Environmental Well-being: AI systems should benefit society and the environment, contributing positively to societal challenges. • Accountability: There should be mechanisms in place to ensure responsibility and accountability for AI systems and their outcomes.

Abbreviation

Abbreviation	AI Act Definition
AI	Artificial Intelligence
EU AI Act	European Union Artificial Intelligence Act
EU	European Union
FLOPs	Floating Point Operations
GPAI	General-Purpose AI
GPAI-M	General-Purpose AI Model
GPAI-M SR	General-Purpose AI Model with Systemic Risk
GPAI-S	General-Purpose AI System
SR	Systemic Risk

Methodological Approach for Classifying GPAI-M as GPAI-M SR

The methodology for classifying "GPAI-M" takes into account four primary components that leads to its classification a "GPAI-M SR". These are outlined below.

Classification Criteria

The classification of GPAI-M SR begins with evaluating specific criteria that identify the potential for significant societal, economic, or technological impacts. These criteria serve as the foundation for determining whether a GPAI model possesses attributes or capabilities that elevate its risk profile.

High-Impact Capabilities

A GPAI model is considered to present systemic risks if it exhibits high-impact capabilities that match or surpass those of the most advanced AI systems available. Such capabilities are defined by the following characteristics:

1. Performance and Versatility:

   Models capable of performing complex, multi-domain tasks with a high degree of accuracy and adaptability are considered high-risk. These capabilities may lead to widespread adoption and integration across critical sectors, increasing the potential for misuse or unintended consequences.

2. Autonomy and Adaptability:

   High-impact models are those that demonstrate significant autonomy and the ability to adapt to new tasks without requiring additional training. This scalability and self-sufficiency enhance their utility but also introduce risks associated with reduced human oversight.

3. Societal and Economic Influence:

   The impact of such models is amplified when they influence vital societal or economic functions. For instance, AI systems integrated into healthcare, finance, or critical infrastructure may pose risks that extend beyond individual users to entire populations.

These capabilities necessitate rigorous evaluation to ensure the responsible deployment of such models.

Computation Thresholds

The cumulative computational resources expended during the training phase of a GPAI model are a key indicator of its potential for systemic risk. Specifically, a model is presumed to have systemic risks if the total computation used during its development exceeds 102510^{25}1025 FLOPs. This threshold provides a quantifiable measure of the model's capacity and complexity, serving as a proxy for its overall capabilities.

1. Threshold Significance:

   The 102510^{25}1025 FLOPs benchmark reflects a critical point where a model's computational foundation likely supports advanced features and functionality. These features may include nuanced language processing, multi-modal input/output capabilities, and sophisticated problem-solving abilities.

2. Indicators of Capability:

   Models meeting or exceeding this threshold are more likely to possess the high-impact capabilities outlined earlier. The threshold serves as a practical guideline for identifying models that warrant closer examination under systemic risk criteria.

3. Evaluation Process:

   Providers must accurately document the computational resources used during training, ensuring transparency and accountability. This documentation allows regulators and evaluators to verify compliance with the threshold requirements and assess the associated risk potential.

Implications for Classification

Models identified through these criteria as having systemic risks require further analysis to determine the full extent of their potential impact. This includes a comprehensive evaluation of their design, deployment, and interaction with users and systems in real-world scenarios.

Providers of GPAI models that meet or exceed the computation threshold must notify regulatory authorities, demonstrating compliance with the classification process. Additionally, they may provide evidence to argue that their model does not pose systemic risks, despite its high computational capacity or advanced capabilities.

By focusing on high-impact capabilities and computation thresholds, the classification criteria establish a robust framework for identifying GPAI models with systemic risks. This ensures that potentially disruptive technologies are subject to necessary safeguards and oversight.

Evaluation Factors

Evaluation factors play a critical role in determining whether a general-purpose AI (GPAI) model presents systemic risks. These factors extend beyond computational thresholds to include a comprehensive analysis of the model's architecture, functionality, and societal impact. This section outlines the key considerations for evaluating the potential risks associated with GPAI models.

Data and Parameters

The quality and size of the training dataset, as well as the number of parameters in the model, are fundamental to assessing its capabilities and risk potential.

1. Dataset Quality:

   The training data's diversity, accuracy, and representativeness directly influence the model's performance and reliability. Poorly curated datasets increase the likelihood of biases, errors, and unintended consequences, particularly when deployed in sensitive domains such as healthcare or finance.

2. Dataset Size:

   Larger datasets typically enable the development of more powerful models, enhancing their ability to perform complex tasks. However, they also introduce challenges related to ensuring that data is ethically sourced and free from inherent biases.

3. **Parameter Count:**

4. The number of parameters within a model determines its complexity and capacity to process nuanced information. While high-parameter models often deliver superior performance, they also present heightened risks of misuse, particularly in applications requiring ethical judgment or social sensitivity.

Computation and Modality

The evaluation of the computation used for training and the model's input/output modalities sheds light on its technical capabilities and adaptability.

1. **Training Computation:**

   The total computation expended during training provides insights into the model's sophistication. Models requiring extensive computational resources tend to exhibit higher versatility and effectiveness but may also possess capabilities that amplify systemic risks.

2. **Modalities:**

   GPAI models can process and generate information across various modalities, such as text, images, and audio. Models that handle multiple modalities seamlessly pose unique risks, as their functionality spans diverse applications, including those with critical societal implications.

Autonomy and Scalability

A model's level of autonomy and scalability is critical for understanding its potential to perform tasks independently and adapt to new contexts.

1. **Autonomy:**

   Autonomous models capable of decision-making without human intervention are particularly concerning in high-stakes environments. The absence of oversight may lead to unintended actions or outcomes, especially if the model encounters situations it was not explicitly trained to handle.

2. **Scalability:**

   Models that scale effectively to address diverse tasks without retraining introduce systemic risks due to their potential misuse. This adaptability, while beneficial for innovation, necessitates stringent monitoring and regulation to mitigate risks.

Market Impact

The model's impact on the internal market is a vital factor in evaluating its systemic risk.

1. **User Base:**

   A GPAI model is presumed to have a significant market impact if it is available to at least 10,000 registered business users within the EU. Such widespread adoption implies that any unintended consequences of its deployment could affect a substantial segment of the population.

2. **Economic Influence:**

   Models integrated into critical sectors such as healthcare, education, and logistics may disrupt existing systems, leading to economic instability or reduced accessibility to essential services.

Comprehensive Risk Assessment

By examining data quality, computational characteristics, autonomy, and market presence, stakeholders can holistically evaluate a model's risk profile. These evaluation factors ensure that decisions about systemic risk designation are based on robust, multi-dimensional analysis, aligning with regulatory requirements and ethical considerations.

Through this approach, the evaluation process identifies GPAI models that may require additional oversight, safeguarding societal and economic interests while fostering technological innovation.

Technical Tools and Methodologies

The classification of general-purpose AI (GPAI) models as having systemic risks requires robust technical tools and methodologies to ensure comprehensive and reliable evaluation. These methodologies, when implemented effectively, provide a structured framework for assessing the potential risks and impacts of GPAI models while mitigating vulnerabilities and enhancing their safety and reliability.

Standardized Protocols for Model Evaluation

Standardized evaluation protocols are integral to assessing systemic risks in GPAI models. These protocols provide consistency and objectivity, ensuring that all evaluations are conducted based on established best practices.

1. Performance Testing:

   Testing the model's performance across a range of tasks is essential to identify its capabilities and potential risks. This includes evaluating its ability to generalize across domains and its limitations in unfamiliar scenarios.

2. Adversarial Testing:

   Adversarial testing involves exposing the model to deliberately challenging inputs designed to exploit its vulnerabilities. This method helps identify weaknesses that may lead to undesirable behaviours, such as biases, errors, or security breaches.

3. Robustness Assessments:

   Robustness testing evaluates how well the model performs under varying conditions, such as noisy data or unexpected inputs. This ensures that the model can maintain reliability and accuracy in real-world applications.

Continuous Monitoring and Risk Assessment

Effective systemic risk evaluation extends beyond the development phase to encompass continuous monitoring during deployment. This approach ensures that the model remains safe and compliant throughout its lifecycle.

1. Real-Time Monitoring:

   Implementing real-time monitoring systems enables stakeholders to track the model's behaviour and detect anomalies or deviations from expected performance. Early detection of potential risks allows for timely mitigation.

2. Incident Documentation:

   Maintaining detailed records of incidents and anomalies provides valuable insights into recurring issues or patterns of risk. Such documentation supports ongoing improvements and informs future regulatory decisions.

3. Dynamic Updates:

   Continuous risk assessment facilitates dynamic updates to the model, addressing newly identified vulnerabilities and incorporating advancements in safety protocols. This iterative process ensures that the model evolves to remain secure and effective.

Tools for Risk Mitigation

The deployment of advanced tools for risk mitigation is critical to reducing the likelihood of systemic failures or adverse impacts.

1. Explainability Tools:

   Enhancing model transparency through explainability tools allows users and evaluators to understand the decision-making processes underlying the model's outputs. This reduces the risk of unintended outcomes and increases trust.

2. Safety Layers:

   Embedding safety mechanisms, such as fallback systems and human oversight controls, ensures that the model operates within predefined safety parameters, even in complex or high-stakes environments.

Collaborative Frameworks

The evaluation and mitigation of systemic risks require collaboration among various stakeholders, including developers, regulators, and independent experts.

1. Regulatory Compliance:

   Alignment with regulatory frameworks ensures that the model adheres to legal and ethical standards, mitigating risks associated with misuse or unintended consequences.

2. Stakeholder Engagement:

   Engaging stakeholders, including civil society and academia, brings diverse perspectives to the evaluation process. This inclusivity ensures that the methodologies are robust and address a broad spectrum of concerns.

Documentation and Reporting

A systematic approach to documentation and reporting is vital for transparency and accountability.

1. Evaluation Reports:

   Comprehensive evaluation reports detail the methodologies used, findings, and recommendations for mitigating identified risks. These reports serve as a reference for stakeholders and regulators.

2. Incident Reporting:

   Immediate reporting of serious incidents to relevant authorities ensures that systemic risks are addressed promptly. This process fosters a culture of accountability and continuous improvement.

The use of standardized protocols, continuous monitoring, and advanced tools, coupled with collaborative frameworks and thorough documentation, ensures a rigorous evaluation of systemic risks in GPAI models. This methodological approach safeguards societal and economic interests while fostering trust and accountability in AI systems.

Reassessment Process

The designation of a general-purpose AI (GPAI) model as having systemic risks is not a static determination. As technology evolves and new insights emerge, a robust reassessment process is essential to ensure the fairness, accuracy, and ongoing relevance of such classifications. The reassessment process is designed to provide flexibility for providers while maintaining rigorous oversight by regulatory authorities. It incorporates mechanisms for addressing changes in model capabilities, usage patterns, and external factors that could impact the risk profile of a model.

Provider-Initiated Reassessment Requests

Providers of GPAI models have the opportunity to request a reassessment if they believe that significant changes or new evidence warrant reconsideration of the model's classification as a systemic risk.

1. Objective Grounds for Reassessment:

   Providers must present clear, objective reasons to justify their reassessment request. These could include:

   - Substantial updates or improvements to the model, such as enhanced safety features or new training methodologies.
   - Changes in the model's deployment context, such as restricted use cases or reduced market availability.
   - New data or evidence that challenges the initial classification criteria, such as improved performance metrics or reduced computational intensity.

2. Timeline for Requests:

   Providers may initiate a reassessment request no earlier than six months after the initial designation decision. This interval ensures that any modifications or new developments are substantial and verifiable.

3. Submission Process:

   Reassessment requests must include detailed documentation outlining the reasons for the request, supporting evidence, and any relevant data. Providers are encouraged to use standardized templates to ensure consistency and facilitate the review process.

Role of the Commission in Reassessment

The Commission, as the regulatory authority, plays a central role in overseeing the reassessment process to maintain transparency, objectivity, and accountability.

1. Evaluation of Reassessment Requests:

   Upon receiving a reassessment request, the Commission will evaluate the submitted documentation against the original classification criteria. This evaluation may involve:

   - Reviewing changes to the model's capabilities, training data, and computational thresholds.
   - Conducting independent testing to verify the claims made by the provider.
   - Consulting with external experts or stakeholders to gather diverse perspectives.

2. Proactive Reassessments:

   In addition to provider-initiated requests, the Commission may independently initiate a reassessment if new information or circumstances arise. This could include:

   - Reports of incidents or adverse impacts related to the model's deployment.
   - Advances in evaluation methodologies that enable more accurate risk assessment.
   - Evidence of significant changes in market dynamics, such as expanded or diminished user adoption.

Outcomes of the Reassessment

The reassessment process may result in several possible outcomes, depending on the findings:

1. Upholding the Classification:

   If the reassessment confirms that the model continues to meet the criteria for systemic risk, the original designation will remain in place.

2. Modification of Classification:

   In cases where evidence supports a change, the classification may be adjusted. For example:

   - A model may be downgraded if new evidence demonstrates reduced risk factors.
   - Conversely, a model could be elevated to a higher risk category if additional concerns emerge.

3. Removal of the Classification:

   If the reassessment determines that the model no longer meets the criteria for systemic risk, the designation may be removed. This decision reflects the Commission's commitment to fairness and evidence-based regulation.

Documentation and Transparency

To ensure transparency, all reassessment outcomes will be documented in detailed reports, which include:

- The rationale for the decision.
- A summary of the evidence and analysis.
- Any recommendations for future monitoring or updates.

These reports will be shared with the provider and, where appropriate, made publicly available to maintain stakeholder trust and accountability.

Continuous Improvement

The reassessment process is an integral component of a dynamic regulatory framework. It ensures that classifications are adaptable to technological advancements and emerging risks, fostering a balanced approach to innovation and safety. By providing mechanisms for re-evaluation, this process upholds the integrity and credibility of systemic risk designations while supporting the responsible development of GPAI models.

Procedure for Classifying GPAI-M as GPAI-M SR

1. Threshold Assessment for High-Impact Capabilities

The threshold assessment for high-impact capabilities is a fundamental procedure in determining whether a GPAI model presents systemic risks. This assessment relies on evaluating the cumulative computational resources utilized during the training of the AI model, with a clear benchmark established at 102510^{25}1025 FLOPs. Below are detailed step-by-step instructions to carry out this process effectively:

Step 1: Establish a Comprehensive Understanding of Computational Metrics

1. Define the Scope of Computation:

   Identify and document all aspects of computation involved in the training process. Include the total training epochs, the hardware configurations (e.g., GPUs, TPUs), and the specific architectures used during training.

2. Calculate Total FLOPs:

   - Determine the number of operations performed per training step.
   - Multiply this by the total number of training steps to calculate cumulative FLOPs.
   - Include computation from auxiliary tasks like pretraining and fine-tuning, ensuring all significant training stages are captured.

Step 2: Compare Computation with the Benchmark

1. Set the Benchmark Threshold:

   Use 102510^{25}1025 FLOPs as the fixed computational threshold, reflecting the level of resource intensity associated with systemic risk.

2. Assess Compliance:

   Compare the calculated total FLOPs of the model against the benchmark. If the total meets or exceeds 102510^{25}1025 FLOPs, the model qualifies for presumed systemic risks.

Step 3: Documentation and Evidence Compilation

1. Compile Technical Documentation:

   Prepare a report detailing the computation analysis, including:

   - Training methodology and hardware.
   - Step-by-step calculation of FLOPs.
   - Evidence of computations performed, such as logs or cloud platform usage records.

2. Ensure Accuracy and Transparency:

   Double-check calculations for accuracy. Use standardized tools or third-party verification services if necessary to validate findings.

Step 4: Notify Relevant Authorities

1. **Notification Requirement:**

   If the model meets or is expected to meet the threshold, notify the AI Office within two weeks. Early notification is mandatory to allow for appropriate regulatory oversight.

2. **Content of Notification:**

   Submit a comprehensive notification that includes:

   - Evidence that the model meets or will meet the computation threshold.
   - Technical details of the model and training process.
   - Any additional information requested by the AI Office.

Step 5: Reassess the Threshold if Necessary

1. **Monitor for Evolving Standards:**

   Stay informed about updates to the threshold as the AI landscape evolves. Adapt training practices accordingly to ensure ongoing compliance.

2. **Document Adaptations:**

   Maintain records of any adjustments made to the training process to accommodate new thresholds or computational practices.

Step 6: Proceed to Risk Mitigation or Reclassification

1. **Presumption of Systemic Risks:**

   If the 102510^{25}1025 FLOPs threshold is surpassed, presume systemic risks unless further mitigating factors are presented and accepted.

2. **Prepare for Subsequent Steps:**

   Models meeting the threshold may require additional evaluation under notification, argument, and classification procedures. Ensure readiness for these next steps by gathering all relevant information in advance.

The threshold assessment for high-impact capabilities is an essential, quantitative procedure that forms the basis for identifying GPAI models with systemic risks. By following these step-by-step instructions, providers can ensure a thorough and transparent evaluation process, aligning with regulatory expectations while maintaining precision in computational assessment.

2. Notification Requirement

The notification requirement is a crucial element in the procedure for identifying GPAI models that may pose systemic risks. It ensures timely communication between providers and regulatory authorities, allowing for appropriate oversight and risk assessment. The following step-by-step instructions outline the obligations and processes for fulfilling notification requirements effectively.

Step 1: Determine the Need for Notification

1. **Identify Threshold Compliance:**

   - Review the model's training computation metrics.

- Confirm if the cumulative computation used for training exceeds 102510^{25}1025 floating-point operations (FLOPs).

2. Preemptive Assessment:
    - Evaluate if there is evidence or plans indicating that the model will meet the 102510^{25}1025 FLOPs threshold in the near future.
3. Triggering Event:
    - Notification must be submitted when the threshold is met or if providers become aware that it will be met.

Step 2: Adhere to the Two-Week Timeline

1. Start the Clock:
    - The two-week notification period begins upon meeting the computation threshold or recognizing that the threshold will be met.
2. Prioritize Early Submission:
    - Avoid delays in submitting the notification. Early compliance reflects proactive risk management and prevents potential penalties.

Step 3: Prepare the Notification

1. Compile the Required Evidence:
    - Collect detailed records demonstrating that the model meets or will meet the computation threshold.
    - Include technical documentation, such as training logs, resource usage reports, and computational benchmarks.
2. Provide Model Details:
    - Summarize key attributes of the AI model, including its purpose, architecture, and input/output modalities.
    - Highlight any unique features or safeguards incorporated into the model.
3. Address AI Office Requirements:
    - Review guidance issued by the AI Office to ensure all necessary data points are included in the submission.

Step 4: Submit the Notification

1. Format the Notification:
    - Follow the AI Office's recommended format for submissions, ensuring clarity and completeness.
    - Include a cover letter outlining the purpose of the notification and summarizing the evidence provided.
2. Deliver the Submission:

- Send the notification through the official channels designated by the AI Office (e.g., secure online portal or physical delivery).
- Retain proof of submission for record-keeping.

Step 5: Optional Provider Arguments

1. **Prepare an Argument if Applicable:**
   - Providers may challenge the presumption of systemic risks by demonstrating specific mitigating factors. For example:
     - Limited application scope (e.g., the model is tailored for non-sensitive tasks).
     - Built-in safeguards to prevent misuse.
2. **Substantiate the Argument:**
   - Provide robust evidence to support claims, such as testing data or third-party validation of safeguards.
3. **Include the Argument with the Notification:**
   - Submit the argument as part of the notification package for simultaneous consideration.

Step 6: Maintain Communication

1. **Respond to Follow-Up Requests:**
   - The AI Office may request additional information or clarification. Ensure prompt responses to expedite the review process.
2. **Track the Review Progress:**
   - Monitor updates from the AI Office regarding the status of the notification and potential next steps.

Step 7: Prepare for Subsequent Classification Steps

1. **Await Feedback from the AI Office:**
   - The notification marks the beginning of a review process that may result in the model's classification as a systemic risk.
2. **Stay Ready for Further Procedures:**
   - Be prepared for additional assessments, mitigation requirements, or reassessments based on the notification outcome.

Meeting the notification requirement is a vital step in the regulatory process for classifying GPAI models with systemic risks. Following this structured approach ensures timely, accurate, and thorough submissions, facilitating compliance and contributing to the responsible deployment of AI technologies.

3. Provider's Arguments

Providers whose AI models meet the threshold for systemic risks based on computational criteria may submit arguments to contest the classification. These arguments must present compelling evidence

that, despite meeting the threshold, the model does not pose systemic risks. Below is a structured approach for preparing, submitting, and supporting such arguments.

Step 1: Assess the Justification for Challenging Classification

1. Review Threshold Determination:
    - Confirm that the model exceeds the 102510^{25}1025 floating-point operations (FLOPs) computation threshold or is projected to do so.
    - Assess whether the model's actual use, design, or safeguards mitigate systemic risks effectively.
2. Identify Supporting Evidence:
    - Collect information to demonstrate why the model does not present systemic risks. Consider factors such as:
        - Narrow application scope.
        - Robust security and ethical safeguards.
        - Limited accessibility or controlled deployment.

Step 2: Gather Comprehensive Evidence

1. Technical Documentation:
    - Provide details on the model's architecture, including its capabilities, training data, and scalability.
    - Emphasize features that constrain high-impact capabilities, if applicable.
2. Risk Mitigation Measures:
    - Detail existing safeguards such as content filtering, restricted outputs, and fail-safes to prevent misuse.
    - Include testing data that validates the effectiveness of these measures.
3. Impact Assessment:
    - Submit evidence showing minimal societal, economic, or environmental impact from the model's operation.
    - If applicable, include third-party audits or certifications validating the model's limited risk profile.

Step 3: Draft a Well-Structured Argument

1. Introduction:
    - Begin with a concise summary of the notification and the reasons for contesting systemic risk classification.
2. Key Claims:
    - Present a structured argument addressing why the model should not be classified as presenting systemic risks, citing specific characteristics such as:

- Restricted functionality.
- Targeted use cases that do not impact critical systems or markets.
- Extensive risk mitigation measures.

3. Evidence Integration:
    - For each claim, attach the supporting evidence gathered, ensuring clarity and coherence in linking the argument to the documentation.
4. Conclusion:
    - Summarize the reasons for contesting classification and request a favorable review based on the evidence presented.

Step 4: Submit the Argument

1. Package the Submission:
    - Include the argument document along with any supporting materials as appendices.
    - Ensure the content is organized and follows the AI Office's submission format guidelines.
2. Deliver Through Official Channels:
    - Submit the argument with the original notification or as an addendum, depending on the timing of the challenge.
3. Confirm Receipt:
    - Request acknowledgment of the submission to ensure it has been received and logged for review.

Step 5: Engage with the Review Process

1. Respond to Requests for Additional Information:
    - The AI Office or Commission may request clarification or supplementary evidence. Respond promptly to maintain the timeline for review.
2. Monitor the Progress:
    - Keep track of communications from the reviewing authority for updates on the argument's assessment.

Step 6: Prepare for Potential Outcomes

1. Approval of Arguments:
    - If the Commission accepts the arguments, the model will not be classified as presenting systemic risks.
    - Maintain documentation of the decision for future reference.
2. Rejection of Arguments:

- If the arguments are deemed insufficient, the model will proceed through the classification process as presenting systemic risks.
- Providers may then focus on compliance with systemic risk obligations or consider filing for reassessment after six months if new evidence arises.

Submitting a well-supported argument against systemic risk classification allows providers to demonstrate that their models, despite meeting computational thresholds, do not pose a significant threat. Following this structured process ensures a clear, evidence-based approach that aligns with regulatory expectations.

4. Commission's Assessment

The Commission's assessment is a critical step in the process of determining whether a general-purpose AI model, after meeting the threshold for high-impact capabilities, should be classified as presenting systemic risks. This section outlines the step-by-step instructions for how the Commission will evaluate the provider's arguments and other relevant factors in making its final classification decision.

Step 1: Review of Provider's Submission

1. Verify Submission Completeness:
    - The Commission will first ensure that all the required information and documentation from the provider have been submitted. This includes evidence demonstrating that the model exceeds the computation threshold (e.g., 10^25 floating-point operations) and any accompanying risk mitigation data.
    - If any necessary information is missing or unclear, the Commission will request additional clarification from the provider.
2. Evaluate the Provider's Argumentation:
    - The Commission will carefully review the provider's argument, especially if the provider claims that the model does not present systemic risks despite meeting the computation threshold. This includes assessing:
        - The scope of the model's intended use and whether it is likely to impact critical societal, economic, or environmental systems.
        - The provider's evidence of risk mitigation measures, such as robust security mechanisms, operational restrictions, or ethical guidelines.
        - Whether these measures are comprehensive and have been effectively tested to limit the model's potential for harm.

Step 2: Independent Risk Evaluation

1. Conduct a Technical Review:
    - The Commission may request or carry out an independent technical evaluation of the model. This may involve:
        - Collaborating with external experts to assess the model's capabilities, risks, and compliance with relevant safety standards.

- Using available protocols, such as adversarial testing or risk modeling, to evaluate potential vulnerabilities and systemic impact that the provider might not have addressed.

2. Consider Impact Analysis:
    - The Commission will evaluate the model's potential risks in terms of its societal, economic, or environmental impact. Key factors include:
        - The extent to which the model could influence market dynamics or public safety, especially in high-stakes areas like healthcare, finance, or infrastructure.
        - Any historical data or case studies regarding similar models and their associated risks.
        - Whether the model's deployment could lead to unintended negative consequences, such as exacerbating inequalities, fostering misinformation, or endangering privacy.

Step 3: Assess the Model's Scalability and Autonomy

1. Scalability Review:
    - The Commission will assess the model's scalability, which refers to its ability to adapt to a wide range of tasks or situations without requiring significant retraining or intervention. If the model can be deployed across different domains or industries, its systemic risks might increase.
    - Considerations include whether the model could be misused in areas where it was not originally intended to operate.
2. Autonomy Evaluation:
    - The Commission will evaluate how autonomous the model is in decision-making. Highly autonomous AI models with the capability to make independent decisions in critical systems pose higher risks. The Commission will assess:
        - The level of human oversight and intervention required during the model's operation.
        - The risks associated with fully autonomous systems in sensitive or high-stakes environments.

Step 4: Decision Making

1. Make a Determination:
    - After reviewing the provider's submission and conducting its own independent analysis, the Commission will decide whether the model presents systemic risks. This determination will be based on the following factors:
        - Whether the model exceeds the computation threshold and presents capabilities that align with high-impact systems.
        - The provider's arguments and supporting evidence regarding risk mitigation.

        - The independent technical assessment and impact evaluation results.
    - If the Commission concludes that the model presents systemic risks, it will proceed with formal classification under applicable regulations.

2. Document and Communicate the Decision:
    - The Commission will formally document its decision, providing clear reasoning behind its classification.
    - The provider will be notified in writing of the Commission's assessment, including any specific compliance requirements or corrective actions that need to be taken.

Step 5: Independent Designation (If Necessary)

1. Initiating Independent Designation:
    - In certain cases, even if the provider does not submit a notification, the Commission may still independently designate a model as presenting systemic risks. This is done using criteria outlined in Annex XIII and can be based on the model's computational characteristics, functionality, or broader societal concerns.
    - The Commission will initiate this process if it believes the model poses an unaddressed systemic risk.

2. Informing the Provider:
    - If the Commission independently designates the model, the provider will be informed, and the appropriate classification will be issued. The provider may then be subject to the associated regulatory obligations, such as compliance with safety measures or adjustments to the model.

The Commission's assessment process is thorough, considering both the provider's arguments and an independent review of the model's potential systemic risks. By following these steps, the Commission ensures that the final decision is based on a comprehensive understanding of the model's capabilities, risks, and societal impact, aligning with regulatory standards and public safety.

5. Reassessment Option

The **Reassessment Option** provides an avenue for AI model providers to request a reevaluation of their model's classification as presenting systemic risks. This reassessment can be requested if new, objective evidence arises after the initial classification. However, the reassessment process is subject to specific conditions and guidelines, which are outlined below.

Step 1: Eligibility for Reassessment Request

1. **Time Frame for Request:**
    - Providers may request a reassessment of their model's classification **only after a minimum period of six months** from the initial designation. This ensures that the model has had sufficient time in the market or operational environment for any changes or new evidence to materialize.

- If the model has been in operation or used widely in that time, there should be enough new information or insights into the model's impact, risks, or performance to warrant a reassessment.

2. Criteria for New Evidence:
   - The reassessment request must be based on **new objective evidence**. This could include:
     - **Technological improvements** or updates to the model that mitigate previously identified risks.
     - **New data** showing that the model's systemic risks are lower than initially estimated.
     - **Changes in market dynamics**, regulatory environment, or use cases that reduce or eliminate the risks the model poses.
     - **Successful risk mitigation strategies** or safety measures that have been implemented, reducing the likelihood of harm.

Step 2: Submitting the Reassessment Request

1. Notification to the AI Office:
   - The provider must formally **notify the AI Office** of their intention to request a reassessment. This must be done through an official communication channel, ensuring the request is logged and tracked appropriately.
   - The request must include a **detailed explanation** of the new objective evidence, including any supporting documents, data, or reports that demonstrate how the model's risks have been reduced or mitigated.
   - If the request involves updates to the model, such as a new version with improved safety measures or performance metrics, the provider must provide documentation of these updates.
2. Supporting Documentation:
   - The provider should provide a comprehensive set of documents to support their reassessment request. This may include:
     - **Technical reports** detailing any updates or changes made to the model's functionality, architecture, or performance.
     - **Risk assessments or impact analysis** showing how the new evidence has altered the model's risk profile.
     - **Market or usage data** showing a reduction in systemic risks or a shift in the model's operational environment that justifies a reassessment.

Step 3: Commission's Review Process

1. Assessment of New Evidence:

- Upon receiving the reassessment request, the Commission will conduct a thorough review of the new evidence submitted by the provider. The key focus will be on:
    - Whether the new evidence significantly reduces or mitigates the identified risks associated with the model.
    - The validity and reliability of the new data or changes, and whether they have been independently verified or assessed by relevant experts.
    - Whether any new risks or unforeseen consequences have emerged as a result of the changes, and whether these pose systemic risks of their own.

2. Engagement with External Experts (if needed):
    - In some cases, the Commission may choose to engage with external experts or consultants to help assess the new evidence. This is particularly relevant if the reassessment involves highly technical changes or if the risks remain complex and difficult to evaluate internally.
    - Expert opinions may include technical assessments, ethical evaluations, or industry-specific reviews of the model's potential impact.

3. Determining the Outcome:
    - After reviewing the reassessment request, the Commission will make one of the following decisions:
        - Model Declassification: If the Commission determines that the new evidence conclusively demonstrates that the model no longer presents systemic risks, it may declassify the model, removing the designation as a high-risk model.
        - No Change in Classification: If the Commission finds that the new evidence is insufficient to significantly reduce the identified risks, the model will remain classified as presenting systemic risks.
        - Revised Classification: In some cases, the Commission may revise the level of risk associated with the model. This could involve reclassifying the model as presenting lesser risks or requiring additional safety measures, but still maintaining a classification of systemic risk.

Step 4: Communicating the Outcome

1. Notification to Provider:
    - Once the Commission has made its decision, it will formally notify the provider of the outcome. The notification will clearly outline the reasons for the decision and provide a detailed explanation of any changes in classification or requirements.
    - If the model is declassified or reclassified, the provider will be informed of any new compliance or regulatory obligations, if applicable.

2. Public Communication (if necessary):
    - If the model's classification is changed, especially if it is declassified, the Commission may choose to make an announcement or update its public records to reflect the new

status. This helps maintain transparency and ensures that other stakeholders, such as regulators or users, are aware of the change.

Step 5: Post-Reassessment Compliance

1. Ongoing Monitoring (if applicable):
    - Even after a reassessment, the Commission may implement **ongoing monitoring** for models that remain classified as presenting systemic risks. This ensures that any future changes, updates, or incidents are captured early, allowing for timely interventions if necessary.
    - If the model is declassified or reclassified, the provider must continue to monitor the model's impact, and should promptly notify the Commission of any significant changes or incidents.

The reassessment option allows AI providers to update the Commission on changes that might reduce their model's systemic risks, offering a mechanism for ongoing flexibility in regulation. The process is designed to ensure that AI models are classified and regulated in a manner that reflects their current capabilities and risks, adapting as needed based on objective, evidence-backed information.

6. Open-Source Model Considerations

The procedure for classifying a general-purpose AI model as presenting systemic risk includes specific considerations for open-source models. These models often pose unique challenges due to their public availability, rapid adoption, and the difficulty in retroactively applying regulatory measures once they are released. To ensure compliance with systemic risk classifications, providers of open-source AI models must adhere to special notification and compliance requirements, outlined below.

Step 1: Early Notification for Open-Source Models

1. Pre-release Notification Requirement:
    - Providers who intend to release a model as open-source must notify the **AI Office before** making the model publicly available. This early notification ensures that the AI Office can assess whether the model meets the criteria for systemic risks prior to its release, thereby preventing complications after the model is freely distributed.
    - The notification must include sufficient documentation to allow the AI Office to evaluate the model's capabilities, including its computation thresholds, impact potential, and any other factors that might present systemic risks.
2. Content of the Notification:
   The notification should include the following key information:
    - **Model details:** A description of the model's architecture, capabilities, and intended use cases.
    - **Training Data Information:** Details on the dataset used for training, including size, scope, and quality of data.
    - **Computation Data:** Evidence that demonstrates whether the model meets or exceeds the threshold of 10^25 floating-point operations, which would indicate potential systemic risks.

- **Risk Assessment:** A preliminary assessment of the potential risks the model might pose once released and any proposed measures for mitigating these risks.

3. **Importance of Early Notification:**
    - Early notification is critical for open-source models to allow the **AI Office** to conduct a risk assessment **before** the model is distributed to the public. This process ensures that all necessary safety, compliance, and regulatory measures are in place. It also allows the provider to address any compliance issues before the model is made available for use, minimizing any post-release complications.

Step 2: Exemptions for Open-Source Models

1. **General Exemption Criteria:**
    - Open-source models are generally **exempt** from certain obligations that apply to proprietary models, especially those that involve detailed compliance requirements, ongoing monitoring, or mandatory notifications to the AI Office.
    - However, these exemptions apply **only if** the open-source model is **not** designated as presenting systemic risks. If the model is deemed to meet the threshold for systemic risk after review, it will be subject to the same regulatory requirements as non-open-source models.

2. **Exemption Limitations:**
    - The exemption for open-source models is limited in scope. Even if a model is released as open-source, it must still comply with specific rules if it is designated as presenting systemic risks. This includes mandatory risk reporting, potential mitigation strategies, and monitoring obligations, just as with proprietary models.
    - Open-source models that pose systemic risks due to their computation capabilities, autonomy, or market impact are subject to **full classification and regulatory processes**, including notifications, compliance checks, and reassessment procedures.

Step 3: Compliance with Regulatory Requirements After Release

1. **Ongoing Responsibility for Risk Mitigation:**
    - Providers of open-source models that have been classified as presenting systemic risks must ensure that they take responsibility for any necessary **risk mitigation** efforts, even after the model is released. This may include providing updates to address newly identified risks or altering the model to better align with regulatory standards.
    - If an open-source model is declassified or reclassified (e.g., no longer presenting systemic risks), the provider must notify the AI Office accordingly and ensure that any public communications reflect the updated status of the model.

2. **Post-release Modifications:**
    - Post-release modifications may be necessary if the AI Office identifies new risks or if the open-source community discovers and reports issues with the model. This is a particular challenge for open-source models since they are often widely distributed and modified by third-party users.

- The provider is expected to actively monitor the model and be responsive to any emerging risks. If systemic risks are identified after release, the provider must work with the AI Office to implement corrective actions.

Step 4: Impact on Open-Source Communities and Users

1. Community and User Responsibility:
    - Open-source models are often adopted by a wide range of users, and those users may also bear responsibility for ensuring that the model is used in ways that align with regulatory requirements.
    - Providers should encourage users to report any issues related to systemic risks or misuse, helping to track the model's impact in the real world. This proactive feedback loop can be crucial for identifying new risks as the model is used in diverse contexts.
2. Informed Use:
    - Providers are encouraged to provide clear guidelines and documentation to help users understand the potential risks associated with the open-source model. This includes highlighting any known limitations, risks, or precautions that should be taken when using the model in production.

Step 5: Ongoing Monitoring and Reporting

1. Continuous Risk Assessment:
    - Even after an open-source model is released, continuous risk monitoring remains essential to ensure that any unforeseen issues are addressed in a timely manner. If the model presents substantial risks, there must be mechanisms for real-time feedback from the user community, regulatory bodies, and stakeholders.
    - Providers of open-source models must remain engaged with regulatory authorities to ensure compliance with any updated risk assessments or mitigation strategies as they arise.
2. Periodic Updates to the AI Office:
    - Open-source model providers must periodically update the AI Office on the model's status, particularly if new evidence arises that significantly alters the model's risk profile. These updates ensure that the AI Office is informed of ongoing developments and can assess whether further intervention is needed.

The release of open-source models presents unique challenges in terms of compliance and systemic risk classification. The early notification requirement ensures that open-source models are assessed for potential risks before they are made widely available, while exemptions provide some flexibility in terms of regulatory burdens. However, if an open-source model is deemed to present systemic risks, it will be subject to the same requirements as proprietary models. Ongoing monitoring and collaboration with the AI Office remain essential to ensuring that risks are adequately managed even after the model is publicly available.

Taxonomy of Systemic Risks

Systemic risks can be drawn from the **First Draft General-Purpose AI Code of Practice[1].** The elements of this taxonomy of systemic risks as a basis for their systemic risk assessment and mitigation.

Types of systemic risks

The following can be treated as systemic risks:

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

Nature of systemic risks

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

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

Sources of systemic risks

Sources of risks, also referred to as "factors of risks" or "drivers of risks", are elements (e.g. events, components, actors and their intentions or activities) that alone or in combination give rise to risks (e.g. model theft or widespread cyber vulnerabilities).

[1] European Commission (2024), 'First Draft of the General-Purpose AI Code of Practice published, written by independent experts', accessible at: https://digital-strategy.ec.europa.eu/en/library/first-draft-general-purpose-ai-code-practice-published-written-independent-experts (last accessed 25th November 2024)

The below can be considered relevant sources of systemic risks:

Dangerous model capabilities

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

- Cyber-offensive capabilities, Chemical, Biological, Radiological and Nuclear (CBRN) capabilities, and weapon acquisition or proliferation capabilities.
- Autonomy, scalability, adaptability to learn new tasks
- Self-replication, self-improvement, and ability to train other models
- Persuasion, manipulation, and deception
- Long-horizon planning, forecasting, and strategising
- Situational awareness

Dangerous model propensities

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

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

Model affordances and socio-technical context

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

- Potential to remove guardrails
- Access to tools (including other models)
- Modalities (including novel and combined modalities)
- Release and distribution strategies

- Human oversight
- Model exfiltration (e.g. model leakage/theft)
- Number of business users and number of end-users
- Offence-defence balance, including the number, capacity, and willingness of bad actors to misuse the model
- Societal vulnerability or adaptation
- Lack of explainability or transparency
- Technology readiness (i.e. how mature a technology is within a given application context) Feedback loops in the use of data, model, and inferences

Annex A – Relevant Provisions of EU AI Act

CHAPTER V

GENERAL-PURPOSE AI MODELS

SECTION 1

Classification rules

Article 51

Classification of general-purpose AI models as general-purpose AI models with systemic risk

1. 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.

2. 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 1025.

3. The Commission shall adopt delegated acts in accordance with Article 97 to amend the thresholds listed in paragraphs 1 and 2 of this Article, as well as to supplement benchmarks and indicators in light of evolving technological developments, such as algorithmic improvements or increased hardware efficiency, when necessary, for these thresholds to reflect the state of the art.

Article 52

Procedure

1. Where a general-purpose AI model meets the condition referred to in Article 51(1), point (a), the relevant provider shall notify the Commission without delay and in any event within two weeks after that requirement is met or it becomes known that it will be met. That notification shall include the information necessary to demonstrate that the relevant requirement has been met. If the Commission becomes aware of a general-purpose AI model presenting systemic risks of which it has not been notified, it may decide to designate it as a model with systemic risk.

2. The provider of a general-purpose AI model that meets the condition referred to in Article 51(1), point (a), may present, with its notification, sufficiently substantiated arguments to demonstrate that, exceptionally, although it meets that requirement, the general-purpose AI model does not present, due to its specific characteristics, systemic risks and therefore should not be classified as a general-purpose AI model with systemic risk.

3. Where the Commission concludes that the arguments submitted pursuant to paragraph 2 are not sufficiently substantiated and the relevant provider was not able to demonstrate that the general-purpose AI model does not present, due to its specific characteristics, systemic risks, it shall reject those

arguments, and the general-purpose AI model shall be considered to be a general-purpose AI model with systemic risk.

4. The Commission may designate a general-purpose AI model as presenting systemic risks, ex officio or following a qualified alert from the scientific panel pursuant to Article 90(1), point (a), on the basis of criteria set out in Annex XIII. The Commission is empowered to adopt delegated acts in accordance with Article 97 in order to amend Annex XIII by specifying and updating the criteria set out in that Annex.

5. Upon a reasoned request of a provider whose model has been designated as a general-purpose AI model with systemic risk pursuant to paragraph 4, the Commission shall take the request into account and may decide to reassess whether the general-purpose AI model can still be considered to present systemic risks on the basis of the criteria set out in Annex XIII. Such a request shall contain objective, detailed and new reasons that have arisen since the designation decision. Providers may request reassessment at the earliest six months after the designation decision. Where the Commission, following its reassessment, decides to maintain the designation as a general-purpose AI model with systemic risk, providers may request reassessment at the earliest six months after that decision.

6. The Commission shall ensure that a list of general-purpose AI models with systemic risk is published and shall keep that list up to date, 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.

Annex B - Frequently Asked Questions

Why do we need rules for general-purpose AI models?

AI promises huge benefits to our economy and society. General-purpose AI models play an important role
in that regard, as they can be used for a variety of tasks and therefore form the basis for a range of downstream AI systems, used in Europe and worldwide.

The AI Act aims to ensure that general-purpose AI models are safe and trustworthy.

To achieve that aim, it is crucial that providers of general-purpose AI models possess a good understanding of their models along the entire AI value chain, both to enable the integration of such models into downstream AI systems and to fulfil their obligations under the AI Act. As explained in more detail below, providers of general-purpose AI models must draw up and provide technical documentation of their models to the AI Office and downstream providers, must put in place a copyright policy, and must publish a training content summary. In addition, providers of general-purpose AI models posing systemic risks, which may be the case either because they are very capable or because they have a significant impact on the internal market for other reasons, must notify the Commission, assess and mitigate systemic risks, perform model evaluations, report serious incidents, and ensure adequate cybersecurity of their models.

In this way, the AI Act contributes to safe and trustworthy innovation in Europe.

What are general-purpose AI models?

The AI Act defines a general-purpose AI model as “an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications” (Article 3(63)).

The Recitals to the AI Act further clarify which models should be deemed to display significant generality and to be capable of performing a wide range of distinct tasks.

According to Recital 98, “whereas the generality of a model could, inter alia, also be determined by a number of parameters, models with at least a billion of parameters and trained with a large amount of data using self-supervision at scale should be considered to display significant generality and to competently perform a wide range of distinctive tasks.”

Recital 99 adds that “large generative AI models are a typical example for a general-purpose AI model, given that they allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks.”

Note that significant generality and ability to competently perform a wide range of distinctive tasks may be achieved by models within a single modality, such as text, audio, images, or video, if the modality is flexible enough. This may also be achieved by models that were developed, fine-tuned, or otherwise modified to be particularly good at a specific task.

The AI Office intends to provide further clarifications on what should be considered a general-purpose AI model, drawing on insights from the Commission’s Joint Research Centre, which is currently working on a scientific research project addressing this and other questions.

What are general-purpose AI models with systemic risk?

Systemic risks are risks of large-scale harm from the most advanced (i.e. state-of-the-art) models at any given point in time or from other models that have an equivalent impact (see Article 3(65)). Such risks can manifest themselves, for example, through the lowering of barriers for chemical or biological weapons development, unintended issues of control over autonomous general-purpose AI models, or harmful discrimination or disinformation at scale (Recital 110). The most advanced models at any given point in time may pose systemic risks, including novel risks, as they are pushing the state of the art. At the same time, some models below the threshold reflecting the state of the art may also pose systemic risks, for example, through reach, scalability, or scaffolding.

Accordingly, the AI Act classifies a general-purpose AI model as a general-purpose AI model with systemic risk if it is one of the most advanced models at that point in time or if it has an equivalent impact (Article 51(1)). Which models are considered general-purpose AI models with systemic risk may change over time, reflecting the evolving state of the art and potential societal adaptation to increasingly advanced models. Currently, general-purpose AI models with systemic risk are developed by a handful of companies, although this may also change over time.

To capture the most advanced models, the AI Act initially lays down a threshold of 10^25 floating-point operations (FLOP) used for training the model (Article 51(1)(a) and (2)). Training a model that meets this threshold is currently estimated to cost tens of millions of Euros (Epoch AI, 2024). The AI Office will continuously monitor technological and industrial developments and the Commission may update the threshold to ensure that it continues to single out the most advanced models as the state of the art evolves by way of delegated act (Article 51(3)). For example, the value of the threshold itself could be adjusted, and/or additional thresholds introduced.

To capture models with an impact equivalent to the most advanced models, the AI Act empowers the Commission to designate additional models as posing systemic risk, based on criteria such as number of users, scalability, or access to tools (Article 51(1)(b), Annex XIII).

The AI Office intends to provide further clarifications on how general-purpose AI models will be classified as general-purpose AI models with systemic risk, drawing on insights from the Commission's Joint Research Centre which is currently working on a scientific research project addressing this and other questions.

What is a provider of a general-purpose AI model?

The AI Act rules on general-purpose AI models apply to providers placing such models on the market in the Union, irrespective of whether those providers are established or located within the Union or in a third country (Article 2(1)(a)).

A provider of a general-purpose AI model means a natural or legal person, public authority, agency or other body that develops a general-purpose AI model or that has such a model developed and places it on the market, whether for payment or free or charge (Article 3(3)).

To place a model on the market means to first make it available on the Union market (Article 3(9)), that is, to supply it for distribution or use on the Union market in the course of a commercial activity, whether in return for payment or free of charge (Article 3(10)). Note that a general-purpose AI model is also considered to be placed on the market if that model's provider integrates the model into its own AI system which is made available on the market or put into service, unless the model is (a) used for purely internal processes that are not essential for providing a product or a service to third parties, (b)

the rights of natural persons are not affected, and (c) the model is not a general-purpose AI model with systemic risk (Recital 97).

What are the obligations for providers of general-purpose AI models?

The obligations for providers of general-purpose AI models apply from 2 August 2025 (Article 113(b)), with special rules for general-purpose AI models placed on the market before that date (Article 111(3)).

Based on Article 53 of the AI Act, providers of general-purpose AI models must document technical information about the model for the purpose of providing that information upon request to the AI Office and national competent authorities (Article 53(1)(a)) and making it available to downstream providers (Article 53(1)(b)). They must also put in place a policy to comply with Union law on copyright and related rights (Article 53(1)(c)) and draw up and make publicly available a sufficiently detailed summary about the content used for training the model (Article 53(1)(d)).

The General-Purpose AI Code of Practice should provide further detail on these obligations in the sections dealing with transparency and copyright.

Based on Article 55 of the AI Act, providers of general-purpose AI models with systemic risk have additional obligations. They must assess and mitigate systemic risks, in particular by performing model evaluations, keeping track of, documenting, and reporting serious incidents, and ensuring adequate cybersecurity protection for the model and its physical infrastructure.

The General-Purpose AI Code of Practice should provide further detail on these obligations in the sections dealing with systemic risk assessment, technical risk mitigation, and governance risk mitigation.

If someone open-sources a model, do they have to comply with the obligations for providers of general-purpose AI models?

The obligations to draw up and provide documentation to the AI Office, national competent authorities, and downstream providers (Article 53(1)(a) and (b)) do not apply if the model is released under a free and open-source license and its parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available. This exemption does not apply to general-purpose AI models with systemic risk (Article 53(2)). Recitals 102 and 103 further clarify what constitutes a free and open-source license and the AI Office intends to provide further clarifications on questions concerning open-sourcing general-purpose AI models.

By contrast, providers of general-purpose AI models with systemic risk must comply with their obligations under the AI Act regardless of whether their models are open-source. After the open-source model release, measures necessary to ensure compliance with the obligations of Articles 53 and 55 may be more difficult to implement (Recital 112). Therefore, providers of general-purpose AI models with systemic risk may need to assess and mitigate systemic risks before releasing their models as open-source.

The General-Purpose AI Code of Practice should provide further detail on what the obligations in Articles 53 and 55 imply for different ways of releasing general-purpose AI models, including open-sourcing.

An important but difficult question underpinning this process is that of finding a balance between pursuing the benefits and mitigating the risks from the open-sourcing of advanced general-purpose AI models: open-sourcing advanced general-purpose AI models may indeed yield significant societal

benefits, including through fostering AI safety research; at the same time, when such models are open-sourced, risk mitigations are more easily circumvented or removed.

Do the obligations for providers of general-purpose AI models apply in the Research & Development phase?

Article 2(8) specifies that the AI Act "does not apply to any research, testing or development activity regarding AI systems or AI models prior to their being placed on the market or put into service."

At the same time, many of the obligations for providers of general-purpose AI models (with and without systemic risk) explicitly or implicitly pertain to the Research & Development phase of models intended for but prior to the placing on the market. For example, this is the case for the obligations for providers to notify the Commission that their general-purpose AI model meets or will meet the training compute threshold (Articles 51 and 52), to document information about training and testing (Article 53), and to assess and mitigate systemic risk (Article 55). In particular, Article 55(1)(b) explicitly specifies that "providers of general-purpose AI models with systemic risk shall assess and mitigate possible systemic risks at Union level, including their sources, that may stem from the development (...) of general-purpose AI models with systemic risk."

In any case, the AI Office expects discussions with providers of general-purpose AI models with systemic risk to start early in the development phase. This is consistent with the obligation for providers of general-purpose AI models that meet the training compute threshold laid down in Article 51(2) to "notify the Commission without delay and in any event within two weeks after that requirement is met or it becomes known that it will be met" (Article 52(1)). Indeed, training of general-purpose AI models takes considerable planning, which includes the upfront allocation of compute resources, and providers of general-purpose AI models are therefore able to know if their model will meet the training compute threshold before the training is complete (Recital 112).

The AI Office intends to provide further clarifications on this question.

If someone fine-tunes or otherwise modifies a model, do they have to comply with the obligations for providers of general-purpose AI models?

General-purpose AI models may be further modified or fine-tuned into new models (Recital 97). Accordingly, downstream entities that fine-tune or otherwise modify an existing general-purpose AI model may become providers of new models. The specific circumstances in which a downstream entity becomes a provider of a new model is a difficult question with potentially large economic implications, as many organisations and individuals fine-tune or otherwise modify general-purpose AI models developed by another entity. The AI Office intends to provide further clarifications on this question.

In the case of a modification or fine-tuning of an existing general-purpose AI model, the obligations for providers of general-purpose AI models in Article 53 should be limited to the modification or fine-tuning, for example, by complementing the already existing technical documentation with information on the modifications (Recital 109). The obligations for providers of general-purpose AI models with systemic risk in Article 55 may be limited in similar ways. The General-Purpose AI Code of Practice could reflect differences between providers that initially develop general-purpose AI models and those that fine-tune or otherwise modify an existing model.

Note that regardless of whether a downstream entity that incorporates a general-purpose AI model into an AI system is deemed to be a provider of the general-purpose AI model, that entity must comply with the relevant AI Act requirements and obligations for AI systems.

What is the General-Purpose AI Code of Practice?

Based on Article 56 of the AI Act, the General-Purpose AI Code of Practice should detail the manner in which providers of general-purpose AI models and of general-purpose AI models with systemic risk may comply with their obligations under the AI Act. The AI Office is facilitating the drawing-up of this Code of Practice.

More precisely, the Code of Practice should detail at least how providers of general-purpose AI models may comply with the obligations laid down in Articles 53 and 55. This means that the Code of Practice can be expected to have two parts: one that applies to providers of all general-purpose AI models (Article 53), and one that applies only to providers of general-purpose AI models with systemic risk (Article 55). Another obligation that may be covered by the Code of Practice is the obligation to notify the Commission for providers of general-purpose AI models that meet or are expected to meet the conditions listed in Article 51 for being classified as a general-purpose AI model with systemic risk (Article 52(1)).

What is not part of the Code of Practice?

The Code of Practice should not address inter alia the following issues: defining key concepts and definitions from the AI Act (such as "general-purpose AI model"), updating the criteria or thresholds for classifying a general-purpose AI model as a general-purpose AI model with systemic risk (Article 51), outlining how the AI Office will enforce the obligations for providers of general-purpose AI models (Chapter IX Section 5), and questions concerning fines, sanctions, and liability.

These issues may instead be addressed through other means (decisions, delegated acts, implementing acts, further communications from the AI Office, etc.).

Nevertheless, the Code of Practice may include commitments by providers of general-purpose AI models to document and report additional information, as well as to involve the AI Office and third parties throughout the entire model lifecycle, in so far as this is considered necessary for providers to effectively comply with their obligations under the AI Act.

Do AI systems play a role in the Code of Practice?

The AI Act distinguishes between AI systems and AI models, imposing requirements for certain AI systems (Chapters II-IV) and obligations for providers of general-purpose AI models (Chapter V). While the provisions of the AI Act concerning AI systems depend on the context of use of the system, the provisions of the AI Act concerning general-purpose AI models apply to the model itself, regardless of what is or will be its ultimate use. The Code of Practice should only pertain to the obligations in the AI Act for providers of general-purpose AI models.

Nevertheless, there are interactions between the two sets of rules, as general-purpose AI models are typically integrated into and form part of AI systems. If a provider of the general-purpose AI model integrates a general-purpose AI model into an AI system, that provider must comply with the obligations for providers of general-purpose AI models and, if the AI system falls within the scope of the AI Act, must comply with the requirements for AI systems. If a downstream provider integrates a general-purpose AI model into an AI system, the provider of the general-purpose AI model must cooperate with the downstream provider of the AI system to ensure that the latter can comply with its obligations under the AI Act if the AI system falls within the scope of the AI Act (for example by providing certain information to the downstream provider).

Given these interactions between models and systems, and between the obligations and requirements for each, an important question underlying the Code of Practice concerns which measures are appropriate at the model layer, and which need to be taken at the system layer instead.

How does the Code of Practice take into account the needs of start-ups?

The Code of Practice should set out its objectives, measures and, as appropriate, key performance indicators (KPIs) to measure the achievement of its objectives. Measures and KPIs related to the obligations applicable to providers of all general-purpose AI models 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 (Recital 109). Moreover, the KPIs related to the obligations applicable to providers of general-purpose AI models with systemic risk should reflect differences in size and capacity between various providers (Article 56(5)), while ensuring that they are proportionate to the risks (Article 56(2)(d)).

When will the Code of Practice be finalised?

After the publication of the first draft of the Code of Practice, it is expected that there will be three more drafting rounds over the coming months. Thirteen Chairs and Vice-Chairs, drawn from diverse backgrounds in computer science, AI governance and law, are responsible for synthesizing submissions from a multi-stakeholder consultation and discussions with the Code of Practice Plenary consisting of around 1000 stakeholders. This iterative process will lead to a final Code of Practice which should reflect the various submissions whilst ensuring a convincing implementation of the legal framework.

What are the legal effects of the Code of Practice?

If approved via implementing act, the Code of Practice obtains general validity, meaning that adherence to the Code of Practice becomes a means to demonstrate compliance with the AI Act. Nevertheless, compliance with the AI Act can also be demonstrated in other ways.

Based on the AI Act, additional legal effects of the Code of Practice are that the AI Office can enforce adherence to the Code of Practice (Article 89(1)) and should take into account commitments made in the Code of Practice when fixing the amount of fines (Article 101(1)).

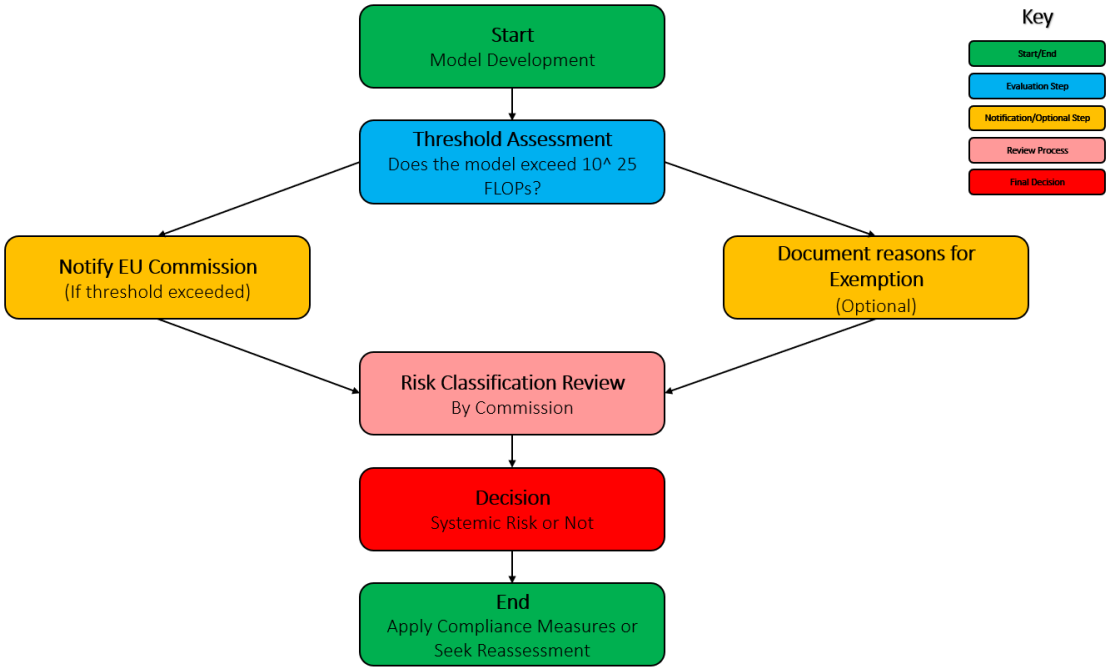
How will the Code of Practice be reviewed and updated?

While the first draft of the Code of Practice does not yet contain details on its review and updating, further iterations of the draft, and any implementing act adopted to approve the final Code of Practice, can be expected to include this information.

Which enforcement powers does the AI Office have?

The AI Office will enforce the obligations for providers of general-purpose AI models (Article 88), as well as support governance bodies within Member States in their enforcement of the requirements for AI systems (Article 75), among other tasks. Enforcement by the AI Office is underpinned by the powers given to it by the AI Act, namely the powers to request information (Article 91), conduct evaluations of general-purpose AI models (Article 92), request measures from providers, including implementing risk mitigations and recalling the model from the market (Article 93), and to impose fines of up to 3% of global annual turnover or 15 million Euros, whichever is higher (Article 101).

Annex C – Workflow Overview for GPAI Classification Process



This workflow provides an overview of the classification process for General-Purpose AI Models under the EU AI Act.

1. **Start with Model Development**: The process begins with the development of a GPAI model.
2. **Threshold Assessment**: Evaluate if the model exceeds 102510^{25}1025 FLOPs in computational resources during training.
3. **Notify the EU Commission** (if the threshold is exceeded) or **Document Exemption** (if seeking exemption due to mitigating factors).
4. **Risk Classification Review**: The Commission assesses the model for systemic risks based on provided data.
5. **Decision**: The model is classified as a systemic risk or not.
6. **Compliance or Reassessment**: Depending on the decision, the provider either applies compliance measures or requests a reassessment if new evidence emerges.

Important notice

This document has been prepared by AI & Partners B.V. for the sole purpose of enabling the parties to whom it is addressed to evaluate the capabilities of AI & Partners B.V. to supply the proposed services.

Other than as stated below, this document and its contents are confidential and prepared solely for your information, and may not be reproduced, redistributed or passed on to any other person in whole or in part. If this document contains details of an arrangement that could result in a tax or National Insurance saving, no such conditions of confidentiality apply to the details of that arrangement (for example, for the purpose of discussion with tax authorities). No other party is entitled to rely on this document for any purpose whatsoever and we accept no liability to any other party who is shown or obtains access to this document.

This document is not an offer and is not intended to be contractually binding. Should this proposal be acceptable to you, and following the conclusion of our internal acceptance procedures, we would be pleased to discuss terms and conditions with you prior to our appointment. Images used throughout the document have either been produced in-house or sourced from publicly available sources

AI & Partners B.V. is the Dutch headquarters of AI & Partners, a global professional services firm. Please see https://www.ai-and-partners.com/ to learn more about us.