# Response by the Centre for Information Policy (CIPL) to the European AI Office's Multi-stakeholder Consultation on Trustworthy General-Purpose AI

Submitted September 18, 2024

The Centre for Information Policy Leadership (CIPL) welcomes the opportunity to contribute to the European Commission's consultation on the Code of Practice for general-purpose AI (GPAI). We encourage the European AI Office to ensure that this Code of Practice is not prescriptive, but rather takes a risk-based approach in applying the proposed practices depending on the context and risk of the GPAI model, as determined by context- or product-specific risk assessments undertaken by covered organisations, consistent with appropriate standards and guardrails reflected in the Code.

We elaborate on this perspective in our narrative responses in the questionnaire, and describe our approach to AI governance and regulation more broadly in these publications:

1. Building Accountable AI Programs: Mapping Best Practices to the CIPL Accountability Framework – https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_building_accountable_ai_programs_23_feb_2024.pdf
2. 10 Recommendations for Global AI Regulation – https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_ten_recommendations_global_ai_regulation_oct2023.pdf

CIPL has elected to leave the selection of boxes in the tables in the questionnaire blank because we find that the provided options for answers would not fully capture how we would approach each of these measures. For example, we do not know how we should interpret the options provided, such as "somewhat agree" and "neither agree nor disagree", and how they would be applied in context. Moreover, the requirement to make single selections in response to each question does not enable due consideration of the contexts and risk levels specific to various models and deployment scenarios.

Below are additional CIPL resources that demonstrate our longstanding history of work relating to AI, particularly at the intersection of AI and data protection. CIPL has produced a range of materials, including white papers, regulatory recommendations, and responses to consultations, that shape the AI governance landscape and bring balanced solutions to existing regulatory and policy challenges. As mentioned in the previous responses, CIPL expertise also centers around accountability, the risk-based approach, and risk management, which are of particular importance to the drafting of this Code.

## Additional CIPL Resources

**White Papers:**

- Automated Decisionmaking and Profiling (ADM) Requirements in U.S. State Privacy Laws, and Current State of Play in State AI Regulations, May 2024
- Enabling Beneficial and Safe Uses of Biometric Technology Through Risk-Based Regulations, April 2024
- Comparison of US State Privacy Laws Data Protection Assessments, February 2024
- Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework, February 2024
- CIPL 10 Recommendations for Global AI Regulation, October 2023

- How the GDPR Regulates AI, March 2020
- Hard Issues and Practical Solutions, February 2020
- Artificial Intelligence and Data Protection in Tension, Oct 2018
- Regulating for Results: Strategies and Priorities for Leadership and Engagement, October 2017
- Recommendations for Implementing Transparency, Consent and Legitimate Interest under the GDPR, May 2017

**Regulatory recommendations, papers, and consultations:**

- CIPL Response to the ICO's 4th Consultation on Engineering Individual Rights into Generative AI Models, June 2024
- CIPL Response to the National Institute of Standards and Technology (NIST)'s Request for Comment on the Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, May 2024
- Response to ICO's 2nd Consultation on Purpose Limitation in the Generative AI Lifecycle, April 2024
- Response to ICO Consultation on the Lawful Basis for Web Scraping to Train Generative AI Models, March 2024
- Response to NIST's Request for Information Related to its Assignments of the Executive Order Concerning AI, February 2024
- CIPL Recommendations on Adopting a Risk-Based Approach to Regulating AI in the EU, March 2021
- CIPL Response to the EU Commission's Consultation on the Draft AI Act, July 2021
- CIPL Response to the EU Commission's AI White Paper, June 2020
- CIPL Examples of Legitimate Interest Grounds for Processing of Personal Data, April 2017

## Section 2. General-purpose AI models with systemic risk: risk taxonomy, assessment and mitigation

### A. Risk taxonomy

**List of Systemic Risks Provided by the Commission:**

Systemic risk from model malfunctions:

- Harmful bias and discrimination: The ways in which models can give rise to harmful bias and discrimination with risks to individuals, communities or societies.
- Misinformation and harming privacy: The dissemination of illegal or false content and facilitation of harming privacy with threats to democratic values and human rights.
- Major accidents: Risks in relation to major accidents and disruptions of critical sectors, that a particular event could lead to a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community.
- Loss of control: Unintended issues of control relating to alignment with human intent, the effects of interaction and tool use, including for example the capacity to control physical systems, 'self-replicating' or training other models.

Systemic risk from malicious use:

- Disinformation: The facilitation of disinformation and manipulation of public opinion with threats to democratic values and human rights.
- Chemical, biological, radiological, and nuclear risks: Dual-use science risks related to ways in which barriers to entry can be lowered, including for weapons development, design acquisition, or use.
- Cyber offence: Risks related to offensive cyber capabilities such as the ways in which vulnerability discovery, exploitation, or operational use can be enabled.

Other systemic risks, with reasonably foreseeable negative effects on:

- public health
- safety
- democratic processes
- public and economic security
- fundamental rights
- the society as a whole.

**Question 10.** Do you consider the following list of systemic risks based on AI Act Recital 110 and international approaches to be comprehensive to inform a taxonomy of systemic risks from general-purpose AI models? If additional risks should be considered in your view, please specify.

| |
|---|
| CIPL cautions against overly broad categories, like "society as a whole" or "safety" and encourages the AI Office to clarify the actual harms (e.g. harassment, extremism) that providers should cover in safety practices. The Office should also clarify which risks are most relevant at each stage of the model lifecycle to ensure that the Code is consistent with the AI Act's risk-based approach in identifying the potential risks of specific uses of AI systems. Overall, it is crucial to address any risks through a holistic, comprehensive approach and support the accountable development and deployment of AI throughout the GPAI lifecycle. |

**Question 11.** What are in your view sources of systemic risks that may stem from the development, the placing on the market, or the use of general-purpose AI models? Systemic risks should be understood to increase with model capabilities and model reach.

Please indicate all relevant elements from the list with an "x". If additional sources should be considered, please specify. You can also provide details on any of the sources or other considerations.

| Indicate Here | Sources of systemic Risks |
|---|---|
| | Level of autonomy of the model: The degree to which a general-purpose AI model has the capability to autonomously interact with the world, plan, and pursue goals. |
| | Adaptability to learn new, distinct tasks: The capability of a model to independently acquire skills for different types of tasks. |
| | Access to tools: A model gaining access to tools, such as databases or web browsers, and other affordances in its environment. |
| | Novel or combined modalities: Modalities a model can process as input and generate as output, such as text, images, video, audio or robotic actions. |
| | Release and distribution strategies: The way a model is released, such as under free and open-source license, or otherwise made available on the market. |
| | Potential to remove guardrails: The ability to bypass or disable pre-defined safety constraints or boundaries set up to ensure a model operates within desired parameters and avoids unintended or harmful outcomes. |
| | Amount of computation used for training the model: Cumulative amount of computation ('compute') used for model training measured in floating point operations as one of the relevant approximations for model capabilities. |
| | Data set used for training the model: Quality or size of the data set used for training the model as a factor influencing model capabilities. |
| **X** | Other (See our response below) |
| | I don't know |

All elements listed could potentially be sources of systemic risks but may not necessarily be so in all instances. It is important to note that GPAI models cannot autonomously interact with their environment unless specifically designed for that purpose. For GPAI models to interact autonomously with their environment, they must be integrated with systems that allow them to do so. Thus, some of these risks may in some circumstances be most relevant to the deployer of the AI model within the context of a system. In any event, the provider should implement transparency and other measures where feasible and appropriate (e.g., model and system cards) to mitigate potential foreseeable risks.

### B. Risk identification and assessment measures

**Question 12.** How can the effective implementation of risk assessment measures reflect differences in size and capacity between various providers such as SMEs and start-ups?

CIPL recognizes the importance of effective risk management by GPAI providers, while considering differences in size, capacity, and resources. The AI Office should take a risk-based approach in their

guidance and provide support (e.g., capacity building) to encourage robust RA practices. The Code should provide guidance on benchmark model performance (e.g., accuracy) and set minimum technical standards that providers must comply with, regardless of maturity level, particularly if their product will be integrated into larger organisations' value chains. Facilitating international alignment is critical to support SMEs, who are disproportionately impacted by divergent regulatory regimes.

**Question 13.** In the current state of the art, which specific risk assessment measures should, in your view, general-purpose AI model providers take to effectively assess systemic risks along the entire model lifecycle, in addition to evaluation and testing?

Please indicate to what extent you agree that providers should take the risk assessment measures from the list. You can add additional measures and provide details on any of the measures, such as what is required for measures to be effective in practice.

| Potential risk assessment measures | Strongly agree | Somewhat agree | Neither agree nor disagree | Disagree | I don't know |
|---|---|---|---|---|---|
| **Determining risk thresholds and risk tolerance**, incl. acceptable levels of risks and capabilities for model development and deployment, and respective quantification of risk severity and probability | | | | | |
| **Forecasting model capabilities and risks** before and during model development | | | | | |
| **Continuous monitoring for emergence of risks,** including data from users, relevant stakeholders, incident databases or similar | | | | | |
| **Determining effectiveness of risk mitigation measures** | | | | | |
| **Safety cases** to demonstrate that the model does not exceed maximum risk thresholds | | | | | |
| **Aggregate risk assessment before model development** | | | | | |
| **Aggregate risk assessment before model deployment** | | | | | |
| **Aggregate risk assessment along the entire model lifecycle** | | | | | |
| **Third-party involvement in risk assessment,** for example, related to inspections of training data, models or internal governance | | | | | |

Providers should adopt risk-specific measures during and after development, where possible. Providers may not have full insight into how models will be used once released and may not have the capacity to foresee and mitigate all risks after development. Providers' ability to monitor and assess risks post-development may differ for closed- vs. open-source models. RA frameworks should consider these factors and be flexible to evolve with rapidly developing technologies and risks. The Commission should provide guidance on RA approaches, including the potential role of voluntary third-party RAs, to foster consistency between organisations.

**Question 14.** Please provide links to relevant material on state-of-the-art risk assessment measures, such as model cards, data sheets, templates or other publications.

- CIPL's report on Building Accountable AI Programs, Appendix A: https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_building_accountable_ai_programs_feb24.pdf
- U.S. Department of State, "Risk Management Profile for AI and Human Rights," https://www.state.gov/risk-management-profile-for-ai-and-human-rights?utm_source=pocket_shared
- MIT FutureTech, "AI Risk Repository," https://airisk.mit.edu/
- IBM, Foundation Models: Opportunities, risks and mitigations, https://www.ibm.com/downloads/cas/E5KE5KRZ

**Question 15.** In the current state of the art, which specific practices related to model evaluations should, in your view, general-purpose AI model providers take with a view to identifying and mitigating systemic risks?

Model evaluations can include various techniques, such as benchmarks and automated tests, red teaming and adversarial testing including stress testing and boundary testing, white-box evaluations with model explanation and interpretability techniques, and sociotechnical evaluations like field testing, user studies or uplift studies.

Please indicate to what extent you agree that providers should implement the practice from the list. You can add additional practices and provide details on any of the practices. You can also indicate which model evaluation techniques listed above or which other techniques can reliably assess which specific systemic risks.

| Potential evaluation practices | Strongly agree | Somewhat agree | Neither agree nor disagree | Disagree | I don't know |
|---|---|---|---|---|---|
| **Performing evaluations at several checkpoints** throughout the model lifecycle, in particular **during development and prior to internal deployment** | | | | | |
| **Performing evaluations at several checkpoints** throughout the model lifecycle, in particular **when the model risk profile changes** such as with access to tools or with different release strategies | | | | | |
| **Ensuring evaluations inform model deployment in real-world conditions** | | | | | |
| **Ensuring evaluations provide insights into the degree to which a model introduces or exacerbates risks** | | | | | |
| **Using non-public model evaluations,** as appropriate | | | | | |
| **Involve independent external evaluators**, including with appropriate | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| levels of access to the model and related information | | | | | |
| **Involve affected persons**, to understand effects of human interactions with a particular model over time | | | | | |
| **Documenting** evaluation strategies and results | | | | | |
| **Reporting** evaluation strategies and results **publicly**, as appropriate | | | | | |
| **Reporting** evaluation strategies and results to **selected authorities and administrative bodies,** as appropriate, including sensitive evaluation results | | | | | |
| **Continuously evaluate and improve evaluation strategies** based on information from risk assessment and mitigation measures, including from incidents and near-misses | | | | | |

Providers may struggle to fully understand how their models are deployed, making certain risk management steps harder to implement. Developers should have flexibility, particularly open-source providers, given these limitations and adopt the following practices where feasible:

- Incorporating feedback loops for responsible user data integration.
- Transparency through public reporting of evaluation strategies, while protecting sensitive information.
- Thorough documentation of model evaluations.
- Adapting RA frameworks to changes in models, data, and application contexts.
- AI/data ethics council to oversee model evaluations, ensuring alignment with industry standards and regulatory requirements.

**Question 16.** Please provide links to relevant material on state-of-the-art model evaluation practices, such as model cards, data sheets, templates or other publications.

- GPT4o System Card (https://openai.com/index/gpt-4o-system-card/) This system card contains details about multi-faceted evaluation practices, including (i) the different stages of Red-teaming, and (ii) the application of OpenAI's safety protocol (Preparedness Framework) to this specific model with details on the risk thresholds and Scorecard.
- Introduction to AI Fact Sheets, IBM (https://aifs360.res.ibm.com/introduction)
- Preparedness Framework (https://cdn.openai.com/openai-preparedness-framework-beta.pdf). This safety protocol details the safety approach OpenAI follows for model development and deployment.
- NIST AI Risk Management Framework (https://www.nist.gov/itl/ai-risk-management-framework)

**Question 17.** What are the greatest challenges that a general-purpose AI model provider could face in implementing risk assessment measures, including model evaluations?

- Distributing responsibility for RA appropriately across actors in the AI value chain.
- Identifying and mitigating risks before deployment, while recognizing the need for ongoing updates as future uses evolve. The difficulty of such predictions makes continuous revisiting and updating of risk assessments vital.
- Conducting thorough real-world scenario testing, balancing resource restraints, and incorporating diverse perspectives to reduce bias.
- Ensuring effective, tailored transparency, while protecting proprietary model information.
- Ensuring appropriate data quality.
- Addressing the lack of global standards and harmonization by leveraging existing frameworks (e.g., NIST AI RMF).

**Question 18.** How can the effective implementation of technical risk mitigation measures reflect differences in size and capacity between various providers such as SMEs and start-ups?

CIPL recommends the Commission encourage implementing effective technical risk mitigation measures across organisations of all sizes. The Commission should consider offering support through accessible resources and capacity-building. Encouraging the sharing of knowledge through benchmarking workshops and fostering industry dialogue may also help smaller entities, such as SMEs and start-ups to implement effective risk mitigation. Providers should also provide transparency about the model's limitations and potential risks so that deployers can implement appropriate safeguards to mitigate those risks and users can make informed decisions in their interactions with the model.

**Question 19.** In the current state of the art, which specific technical risk mitigation measures should, in your view, general-purpose AI model providers take to effectively mitigate systemic risks along the entire model lifecycle, in addition to cybersecurity protection?

Please indicate to what extent you agree that providers should take the measures from the list. You can add additional measures and provide details on any of the measures, such as what is required for measures to be effective in practice. [*Please see comments in text box below*]

| **Potential technical risk assessment measures** | **Strongly agree** | **Somewhat agree** | **Neither agree nor disagree** | **Disagree** | **I don't know** |
|---|---|---|---|---|---|
| **Data governance** such as data selection, cleaning, quality control | | | | | |
| **Model design and development to achieve an appropriate level of trustworthiness characteristics** such as model reliability, fairness or security | | | | | |
| **Fine-tuning for trustworthiness** and alignment such as through Reinforcement Learning from Human Feedback (RLHF) or Constitutional AI | | | | | |
| **Unlearning techniques** such as to remove specific harmful capabilities from models | | | | | |
| **Technical deployment guardrails,** such as content and other filters, capability restrictions, fine-tuning restrictions or monitoring-based | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| restrictions in case of misuse by users | | | | | |
| **Mitigation measures relating to the model architecture, components, access to tools or model autonomy** | | | | | |
| **Detection, labelling and other measures related to AI-generated or manipulated content** | | | | | |
| **Regular model updates**, including security updates | | | | | |
| **Measuring model performance** on an ongoing basis | | | | | |
| **Identification and mitigation of model misuse** | | | | | |
| **Access control to tools and levels of model autonomy** | | | | | |

*Please specify, including the extent you agree that providers should take the measures from the list:*

Model providers should adopt strong data governance and embed trustworthiness into model design from the early stages. Regularly measuring performance, updating, and fine-tuning models to ensure they function as intended can help guide risk mitigation efforts. Deployment guardrails, unlearning techniques, and measures for AI-generated content are also key to managing harmful outputs. On the input side, CIPL also cautions against the systematic removal data of from certain groups by excluding potentially necessary data for bias mitigation.

CIPL urges the Commission to clarify the responsibilities of model providers and deployers, recognizing that the appropriate allocation of these responsibilities will often depend upon the specific context and circumstances of deployment.

**Question 20.** Please provide links to relevant material on state-of-the-art technical risk mitigation practices, such as model cards, data sheets, templates or other publications.

- Model Cards for Model Reporting, Mitchell, Wu, et al., October 5, 2018, available at https://doi.org/10.48550/arXiv.1810.03993
- National Institute for Standards and Technology, "Assessing Risks and Impact of AI," https://ai-challenges.nist.gov/aria.

**Question 21.** What are the greatest challenges that a general-purpose AI provider could face in implementing technical risk mitigation measures?

Model providers face numerous challenges in implementing technical risk mitigation measures, including:

- Managing risk mitigation across GPAI models and integrating trustworthiness into model design as GPAI models can serve different purposes and applications.
- Continuous monitoring/updating of the model demands substantial resources – potentially burdensome for SMEs and may hinder innovation.
- Identifying/managing emergent capabilities of GPAI – AI providers may not have complete visibility/control over all outcomes and risks.

- Maintaining robust data governance (e.g., quality, accuracy, ethical sourcing of data).
- Ensuring consistency with fast-evolving regulatory requirements.

**Question 22.** How can the effective implementation of internal risk management and governance measures reflect differences in size and capacity between various providers such as SMEs and start-ups?

As mentioned previously, all organisations must ensure the effective implementation of internal risk management and governance measures, regardless of their size and capacity. CIPL urges the Commission to take a risk-based, proportional approach by offering support through accessible resources and capacity building initiatives, such as workshops. Shared tools, guidance, and templates for RA frameworks can also enable smaller providers to manage risks effectively while still encouraging innovation of GPAI technology.

**Question 23.** In the current state of the art, which specific internal risk management and governance measures should, in your view, general-purpose AI model providers take to effectively mitigate systemic risks along the entire model lifecycle, in addition to serious incident reporting?

Please indicate to what extent you agree that providers should take the measures from the list. You can add additional measures and provide details on any of the measures, such as what is required for measures to be effective in practice.

| **Potential internal risk assessment and governance measures** | **Strongly agree** | **Somewhat agree** | **Neither agree nor disagree** | **Disagree** | **I don't know** |
|---|---|---|---|---|---|
| **Risk management framework across the model lifecycle** | | | | | |
| **Internal independent oversight functions** in a transparent governance structure, such as related to risks and ethics | | | | | |
| **Traceability** in relation to datasets, processes, and decisions made during model development | | | | | |
| **Ensuring that staff are familiar with their duties** and the organisation's risk management practices | | | | | |
| **Responsible scaling policies** | | | | | |
| **Acceptable use policies** | | | | | |
| **Whistleblower protections** | | | | | |
| **Internal resource allocation towards risk assessment and mitigation** measures as well as research to mitigate systemic risks | | | | | |
| **Robust security controls** including physical security, cyber security and information security | | | | | |
| **External accountability measures** such as third-party audits, model or | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| aggregated data access for researchers | | | | | |
| **Other collaborations and involvements of a diverse set of stakeholders**, including impacted communities | | | | | |
| **Responsible release practices** including staged release, particularly before open-sourcing a model with systemic risk | | | | | |
| **Transparency** reports such as model cards, system cards or data sheets | | | | | |
| **Human oversight** mechanisms | | | | | |
| **Know-your-customer** practices | | | | | |
| **Logging, reporting and follow-up of near-misses** along the lifecycle | | | | | |
| **Measures to mitigate and remediate deployment issues** and vulnerabilities | | | | | |
| **Complaints handling and redress mechanisms, such as bug bounty programs** | | | | | |
| **Mandatory model updating policies and limit on maximum model availability** | | | | | |
| **Third-party and user discovery mechanisms and reporting** related to deployment issues and vulnerabilities | | | | | |

The table above is difficult to complete in the abstract, as different practices may be more or less applicable in specific contexts, e.g., for open- vs. closed-source models. Furthermore, whether to employ and/or require third-party mechanisms depends on the context and potential for risks of the GPAI model.

**Question 24.** Please provide links to relevant material on state-of-the-art governance risk mitigation practices, such as model cards, data sheets, templates or other publications.

- CIPL Report on "Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework" - https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_building_accountable_ai_programs_23_feb_2024.pdf
- Singapore PDPC's Model AI Governance Framework - https://www.pdpc.gov.sg/help-and-resources/2020/01/second-edition-of-model-artificial-intelligence-governance-framework
- IBM's AI ethics governance framework serves as an effective example of how establishing an internal governance framework can help streamline the process of identifying and managing ethics concerns arising from AI projects - https://www.ibm.com/blog/a-look-into-ibms-ai-ethics-governance-framework/

**Question 25.** What are the greatest challenges that a general-purpose AI provider could face in implementing governance risk mitigation measures?

- Effective oversight and governance for comprehensive monitoring and risk mitigation, especially for risks post-deployment.
- Navigating diverse regulatory requirements across jurisdictions, particularly for entities with limited
- resources.
- Ongoing engagement and understanding of the technologies among employees and sharing the responsibility of risk management across the company.
- Engaging with external stakeholders to incorporate diverse perspectives.
- Transparency measures that are both sufficiently effective & protective of proprietary information.
- Security of GPAI model against potential risks (e.g., data breaches, adversarial attacks).
- Scarcity of human experts in technical aspects of AI and data science.