B·M·G
AVOCATS

Newsletter
Geneva, 29 January 2025
Author : Prof. Philippe Gilliéron, philippe.gillieron@bmglaw.ch

AI Governance – the Code of Practice as a road towards compliance and trust?

On December 19, 2024, the second draft of the Code of Practice meant to detail the requirements expected by general-purpose AI providers as set out in Art. 51 *et seq*. of the EU AI Act was published.

This article attempts to provide a snapshot of this 65 pages document.

I. The context: art. 56 of the EU AI Act

In accordance with Art. 56 of the EU AI Act, the AI Office shall encourage and facilitate the drawing of Codes of Practice to enable the providers of general-purpose AI models to more easily comply with their obligations set out in Art. 53 *et seq*. of the EU AI Act, as expressly stated in Art. 56 (4) of the EU AI Act.

Bearing in mind that the provisions of the EU AI Act related to the providers of general purpose AI models will come into force on August 2, 2025, the Codes (which actually rather is "a" code) will have to be assessed by the AI Office and the Board by May 2, 2025 (art. 56 (6) and (9)).

While the Commission may, by way of implementing act, approve a code and give it a general validity within the UE (art. 56 (6)), the code is not meant to be mandatory as "*the AI Office may invite all providers of general-purpose AI models to adhere to the codes of practice*" (art. 56 (7)).

Truth however is that such adherence will make it easier for these providers to demonstrate compliance and will thus make sense, although a three months term to ensure compliance may sound fairly short for certain categories of providers. Without going here into the details of these requirements, suffice it to say that the EU AI Act makes a distinction between the following providers and related transparency requirements:

Nature of the model	Technical doc. (training, testing, evaluation, Annex XI)	Doc. to providers for integration (downstream, Annex XII)	Copyright policy	Summary about the content used for training	Model Evaluation	Risk assessment	Incident	Cyber
Open source			X	X				
Proprietary	X	X	X	X				
Systemic risk (including open source)	X	X	X	X	X	X	X	X

The codes of practice notably have to provide:

(1) the means to ensure that the technical documentation and the ones provided to downstream providers is kept up to date;

(2) an adequate level of details for the summary about the content used for training;

(3) the identification of the type and nature of the systemic risks including their sources; as well as

(4) the measures, procedures and modalities for the assessment and management of these systemic risks.

To ensure that providers comply with these obligations, each of these categories has to be related to KPIs.

II. The structure of the second draft of the AI Code of practice

Four working groups have been appointed to focus on the following respective areas:

(1) Transparency and copyright-related rules;
(2) Risk assessment for systemic risk;
(3) Technical risk mitigation for systemic risk;
(4) Governance risk mitigation for systemic risk.

The first part of the draft Code details transparency and copyright obligations for all providers of general-purpose AI models, with notable exemptions for providers of certain open-source models. The second part, which is the lengthiest part of the draft, is only relevant for a small number of providers of most advanced general-purpose AI models that could pose systemic risks, as defined in art. 51 of the EU AI Act.

On November 27, 2024, a first draft was published. Following a period of consultation, the second draft was released on December 19, 2024. A third version should normally be expected by February 17, 2025.

The second is split into 21 commitments, related measures and KPIs. These can tentatively be summed up in a fairly concise but – hopefully – accurate way in the following table:

III. The content of the second draft of the AI Code of Practice

Areas	**Commitments**	**Content**	**Note**
Transparency	1. Documentation (annexes XI and XII)	General information	
		Intended tasks, type and nature of AI systems for integration (notably high risk)	Ensure that intended and unintended uses of the models are made clear for downstream providers
		Acceptable use policies	
		Date or release and methods of distribution	
		Interaction of the model with external hardware or software	Enable downstream providers to have basic understanding to be used

		Versions of relevant software applicable		Required software to integrate the model into systems
		Architecture and number of parameters		
		Modality and format of inputs/outputs		
		License with the released assets (data, weights, source code)		
		Technical means for integration into AI systems		
		Design specifications of the model and training process		
		Information on data used for training, testing and validation		Proivde information s to data sourcing, processing and overall properties
		Computational resources		Provides information on computational requirements of model training and inference
		Known or estimated consumption		
		Testing process and results		
Copyright	2. Copyright policy (art. 53(1)(c))	**Measures**	**KPI**	
		Draw and implement	(1) single document; (2) versioning; (3) person in charge	
		Publish a summary		
		Assess copyright compliance of third party datasets	(1) Document assurances obtained (2) Document providers' copyright compliance assessment (not SMEs)	Due diligence (copyright status, lawful access, compliance with Art. 3 and 4 Directive 2019/790); reasonable efforts
		Ensure lawful access to copyrighted-protected content	Document measures taken by providers to ensure lawful access (not SMEs)	
		Not crawl websites making available copyright-infringing content	Document list of piracy websites excluded (not SMEs)	
		Respect robot exclusion protocol	Name any new or modified crawler as well as their purpose	State of art technologies to comply with Art. 4(3) Directive 2019/790
		Identify and comply with other appropriate	(1) List other solutions for	

		expressions of rights reservation	expressions of rights reservations; (2) Document attendance to standard-setting meetings (not SMEs)	
		Publish information on rights reservation compliance	Information about measures adopted made publicly available	
		Prevent copyright-related overfitting	Document measures taken to avoid overfitting (not SMEs)	Mitigate the risk that downstream AI sytems generates copyright infringing output identical to procted works
		Prohibit copyright infringing uses	AUP, T&C	
		Designate point of contact	(1) Designation and related information publicly available (2) internal process to handle copyright complaints made publicly available (not SMEs)	
Systemic Risks – general commitments (art. 55(1))	3. Taxonomy	Identification		
		Selected systemic risks		Cyber offense; chemical, biological, nuclear, large-scale manipulation, large-scale discrimination, loss of human oversight
		Sources		Capabilities, propensities, deployment
	4. Safety/security framework (art. 55(1))	Procedures	Percentage of commitments 6-21 having procedures in place (target 100%)	
		Risk tiers	Percentage of identified risks with unacceptable risk tier (target 100%)	Each identified risk includes a tier at which level is considered unacceptable
		Mapping risk tiers to mitigations	Percentage of risk tiers that have described mitigations (target 100%)	

		Forecasting	Percentage of risk tiers not yet reached that have best efforts estimates of timeline (target 100%)	Best efforts of timelines when they expect to develop model that complies with each risk tier
		Serious incident reporting		Processes to be in place with pre-defined corrective measures
	5. Safety/security model reports	Level of detail		Model report proportionate to the highest risk tier the model reaches
		Results of risk ass. and mitigation	(1) Percentage of identified systemic risk for which the Model identifies level with and without safety/security mitigations (target 100%) (2) Percentage of documented deviations from rigour in risk assessment and impact	
		Reasoning for deployment decisions		See commitment 13. Report should also state conditions under which reasoning would no longer hold
		External assessment		
		Algorithmic improvement		
		Technical documentation		
	6. Risk ass. and mitigation along lifecycle	Before training		Risk management process (identification, analysis, evaluation, mitigation, residual risk) along the whole lifecycle should be in place
		During training		
		Before deployment		
		During deployment		
		After retirement		
Systemic Risks – risk assessment	7. Risk identification			
	8. Risk analysis	Methodologies		
		Risk estimation		Will include those in commitment 10 and be built upon serious incident

				reporting (commitment 17)
	9. Risk evaluation			Compare the results of risks analysis (commitment 8) to pre-defined risks (commitment 3)
	10. Evidence collection	Model-independent evidence	Average number of model-independent forms of evidence for the highest risk tier used in a Model Report (target 2)	
		State-of-the-art model evaluatiom	(1) Percentage of evaluations for highest risk tier considered state-of-the-art (target 80%) (2) Percentage of evaluations for lowest risk tier considered state-of-the-art (target 50%)	Methods accepted by a majority of AI safety researchers to be among the best. Every 6 months
		Rigorous model evaluations	(1) Percentage of model evaluations used as evidence which show high internal and external validity (target 75%) (2) Percentage of model evaluations used as evidence for highest risk tier which have been reproduced (target 50%) (3) Percentage of model evaluations used as evidence which a fully portable (target for non-SMEs 25%; for SMEs: 0%) (4) Percentage of model evaluations used as evidence where uncertainty is reported (target 100%)	
		Model elicitation	(1) Average percentage of hours spent on model elicitation for	Can include fine-tuning, prompt engineering, scaffolding, ability

			evaluations used as evidence for highest risk tier (target 75%) (4) Average percentage of hours spent on model elicitation for evaluations used as evidence for highest risk tier (target 33%)	to disable safeguards, etc. Model elicitation with increased risk profile are matched with increased security measures
		Model as part of systems		Ensure that model evaluations can assess capabilities and limitations when integrated into AI systems. Ensure in licensing terms that licensee has to evaluate the model when integrated into the contemplated AI system
		Representative model evaluations & generalisations		
		Exploratory work	Percentage of internal staff hours spent on safety research as part of all research hours (target non-SMEs: 10%; target SMEs and open source models: 0%	
		Sharing tools & best practices	(1) Number of relevant parties with whom evidence collection best practice guidance has been shared (target non-SMEs 10; target SMEs 0%) (2) Number of new model evaluation tools shared with at least 5 providers or evaluators (target non-SMEs: 3; target SMEs and open-source models: 0)	
		Qualified model evaluators and evaluation access		
		Safety margin		

		Forecast-ready model evaluations	Percentage of evaluations used as evidence which allow for some level of forecasting (target 10%)	
		Post-deployment monitoring	(1) Percentage of risks at highest risk tiers for which evidence is being collected using at least 3 forms of post-deployment monitoring (target 75%) (2) Percentage of risks at lowest risk tiers for which evidence is being collected using at least one form of post-deployment monitoring (target 50%) (3) Percentage of licensees who report actionable evidence collected for highest risk tiers (target 50%)	Methods may include anonymous reporting channels, incident report forms and bug bounties, community-driven evaluations, evidence of model usage in the real world, scientific studies of the model, etc.
Systemic Risks – technical risk mitigation (art. 55(1)(b) and (d))	11. Safety mitigations	Mitigations to consider		(a) filtering and cleaning training data, (b) monitoring and filtering the input and outputs of the models, (c) changing the behaviour of the model (fine-tuning to refuse certain requests), (d) restricting access to vetted users, (e) safety tools made available to others to reduce systemic risk, (f) high-assurance quantitative safety guarantees on model behaviour
		State-of-the-art		
	12. Security mitigations	General cybersecurity best practices		(a) strong password policy and mgmt.; (b) email filtering; (c) protection of wireless networks; (d) policies for untrusted

				removable media; (e) physical intrusion prevention; (f) software updates and patch management. See ISO/IEC 27001, NIST 800-53 as well as NIS2 Directive and Cyber Resilience Act.
		Security assurance		(a) frequent security review by accredited third party; (b) frequent active red-teaming; (c) secure communication channels for third parties to report issues; (d) competitive bug bounty programs to encourage public participation; (e) whistleblower policies which prohibit retribution (see commitment 18); (f) install EDR and IDS. See NIST 800-53
		Protection of stored model weights and related assets		(a) registry of all devices and location where model weights are stored; (b) access control and monitoring on such devices; (c) storage on dedicated devices; (d) ensure model weights always encrypted in storage and in transit; (e) decryption to non-persistent memory; (f) restricting access to data centres. See NIST 800-53
		Interfaces and access control to model weights	Percentage of defined interfaces and access point to unreleased model weights and associated assets that have implemented the specified security controls (target 100%)	See NIST 800-53

		Insider threats		Background checks and training to recognize and report insider threats. See NIST 800-53
		Regime of applicability	Percentage of copies and versions of unreleased model weights that have the above security measures in place (target 100%)	All the above should apply to all versions and copies
		Limited disclosure		Only publicly disclose security mitigations measures that do not undermine their effectiveness
	13. Dev. and deploy. Decisions	Conditions for not proceeding		
		Decision process for proceeding		
		Staging deployment when proceeding		(a) limiting API access to vetted users; (b) gradually expanding access based on risk assessments; (c) starting with closed release before open release; (d) logging systems to track usage and safety concerns; (e) criteria for progressing through stages; (f) ability to restrict access.
		Transparency into external input in decision-making		
Systemic Risks – governance risk mitigation (art. 55(1)(b))	14. Systemic risk respons. Allocation		(1) Mgmt body in its mgmt. function: for SMEs, specific individual in executive team has responsibility for risk management efforts (sufficient staff and time); for non-SMEs, Chief Risk Office appointed (with well-staffed risk function) (2) Mgmt body in its supervisory	

			function: specific individuals (such as in audit or external report) oversee organisation's mgmt. of systemic risk; access to information they need. (3) adherence and adequacy assessment (commitment 15) concludes that appropriate levels of responsibility and resources have been allocated.	
	15. Framework adherence and adequacy assessment		(1) Conducted every 6 months; (2) Positive results; (3) Concerns addressed before model is placed on the market or the next scheduled assessment (whatever comes first) (4) External auditor provides positive assessment (4) Framework comparable in rigor, breadth and depth to the ones used by others providers of models with systemic risks.	Thoroughness of the review to be proportional with the possible systemic risks posed by the model in question
	16. External risk assessment	Before market placement After market placement	(1) secure and non-restrictive access protocols are documented and published for external assessors; (2) Bug bounty programs with clear success criteria, financial or compute-based rewards are established; (3) Access to deployed models is granted to external assessors within 30 days upon formal request;	To release a model in open source is a way to provide such access. If only weights are open, research on the model itself should be facilitated

			(4) Policy in place to enable external assessors to publish findings after 60 day notification period; (5) Independent test results are included in the Model Report; (6) safe harbour policy adopted to protect external assessors from legal or financial consequences.	
	17. Serious incident reporting (art. 55(1)(c))			Following should be reported (but no KPI): (a) start and end date of incident; (b) nature and resulting harm; (c) description including chain of events that led to it; (d) root cause analysis (including description of model's outputs and factors that led to it, such as inputs or circumvention of safeguards); (e) model potentially involved in the incident.
	18. Whistleblowing protection (art. 87)		(1) documented evidence of a policy protecting employees; (2) description of communication strategy for informing employees about protection and reporting channels; (3) Evidence of regular awareness initiatives conducted at least annually; (4) Anonymous internal reporting mechanisms (compliance hotline, secure message platform);	See Directive 2019/1937

			(5) Documentation of reports received (confidential); (6) Documentation of initial response time and time taken to react upon the raised concern.	
	19. Notifications (art. 52(1) and 55(1))	General-purpose AI model with systemic risk notification		Before training, estimate computational power to be used and inform AI Office within 2 weeks if above threshold of Art. 51(2)
		Framework update notification		Public link to enable AI Office to have access to unredacted latest version of framework within 5 days of update.
		Framework adequacy assessment notification		Same as above
		Safety and security Model Report notification		
	20. Documentation (art. 53(1)(a) and 55(1))	Re: classification based on Art. 51		Includes documentation required in commitment 1 and results from risk assessments in line with commitments 6-10 as well as number of registered business users established in the EU
		Re: adherence to the Code and the AI Act		Includes Framework, Model Report, Adequacy and adherence assessments (commitment 15), a detailed description of the system architecture and evidence of whistleblower policy and related information to employees.
	21. Public transparency		(1) Publish latest version of Framework within	

			15 business days of sending it to AI Office; (2) Publish Model Reports within 5 days after the relevant model is released in the EU, including a summary.	

IV. Comments

While the final version is yet to come, the commitments, measures and KPIs suggested do not only deserve the full attention of general-purpose AI model providers, but also provide valuable insights for any company in terms of AI governance.

The following points are notably to be kept in mind on that regard for any AI related project:

- Commitment 1 may provide guidance in relation of the points to be taken into account when one intends to develop a downstream AI system (or model obviously), notably in relation to data governance (art. 10 EU AI Act), but also in relation to the assessment of the underlying model (or third party products to be used), its capabilities, propensities or limitations.
- Commitment 2 provides guidance as to the way risks of copyright infringement may be mitigated, *i.e* (1) an assessment of compliance of third party datasets, (2) ensuring lawful access to copyrighted protected content, (3) not crawl websites (obviously) making copyright infringing material available; (4) respect robot exclusion protocol and other reservation of rights (notably in light of Art. 4(3) Directive 2019/790), (5) prevent copyright overfitting, (6) prohibit copyright infringing uses in an acceptable use policy and (6) document all the above in a copyright policy.
- Commitments 3-16 will assist companies in setting up adequate risk management processes by addressing the right questions in the right order (notably in light of Art. 9 of the EU AI Act where applicable), *i.e.*: (1) identification of the nature and categories of risks; (2) classification of risks in different tiers; (3) risk assessment along the AI lifecycle (before training, during training, before deployment, during deployment, after retirement); (4) evidence collection and its sources; (5) phases of the risk assessment (along the NIST framework) and mitigation measures; (6) decision to deploy or not; (7) safety and security measures; (8) internal allocation of responsibilities.
- Commitment 17, although fairly high level, draws attention upon the importance to have incident reporting processes in place.

As a result, the Code of Practice will not only be a MUST be for general-purpose AI providers, but will also provide useful guidance for any company in its efforts to implement proper AI governance processes. It remains to be seen what the final version of the Code will look like.