SaferAI

An Opinionated Literature Review to Inform the EU Codes of Practice on GPAI with Systemic Risks

14 June 2024

simeon@safer-ai.org

An Opinionated Literature Review to Inform the EU Codes of Practice on GPAI with Systemic Risks

The obligations imposed by the EU AI Act on GPAI with Systemic Risks crucially rely on implementing and maintaining an adequate AI risk management system. Therefore, we're proposing in this piece an opinionated literature review focused on risk management that can inform the EU Codes of Practice.

Risk Management

Risk management is a literature that has grown over decades thanks to the accumulated experience across numerous industries. As a result, the risk management process has been harmonized across industries. Raz & Hillson's (2005) comprehensive review of existing risk management practices reveals five common steps shared by almost all of these processes:

1. Planning - Establishing the context, resource allocation, risk thresholds, governance, risk management policy, roles and responsibilities, process, criteria, etc.
2. Identification - Identifying potential risks and risk sources.
3. Analysis - Estimating the probability and consequences of identified risks, and evaluating and prioritizing risks based on defined criteria.
4. Treatment - Identifying and implementing appropriate risk treatment options and plans for the prioritized risks. Common options include avoiding, reducing probability, limiting consequences, and transferring risk.
5. Control & Monitoring - Monitoring and reviewing the performance and effectiveness of the risk management process and the status of identified risks and treatment actions. Making modifications as needed.

In recent years, a growing number of papers and publications has begun to address the details of applying these steps to AI systems. While some aspects still require further research and clarification, significant progress has been made in fleshing out the specifics of each step. In the following sections, we will provide an overview of the existing literature on this topic.

Planning

As part of the planning process, a number of organizational structures and environmental variables matter. The presence of an internal audit function (Schuett, 2023), the decision power of an AI ethics board (Schuett, Reuel & Carlier, 2023) or the implementation of the tried and tested 3 Lines of Defense model (Schuett, 2022) are all indicators of the maturity of an organization's governance structure. These factors should be taken into account

by organizations that are trying to adequately manage systemic risks associated with AI systems. The presence of such functions also reflects a more general attitude towards risks and safety which is generally captured by the notion of safety culture. Manheim (2023) reflects on how organizations developing advanced AI systems may be able to learn from other industries to improve their safety culture, which will downstream have numerous positive effects on the management of systemic risks.

Identification

The identification of systemic risks in GPAI models will require a combination of methods, ranging from standard risk identification techniques to red-teaming and measurement procedures which we'll cover in "Analysis". Koessler et al. (2023) provide a comprehensive overview of how traditional risk assessment techniques can be applied to advanced AI systems. They recommend that providers of GPAI with systemic risks consider risk identification methods such as scenario analysis, the fishbone method, and risk taxonomies and typologies. We suggest papers using risk identification methods we consider most relevant to GPAI with systemic risks: Weidinger et al. (2021) propose a taxonomy covering all risks, with a low level of granularity, while Hendrycks et al. (2023) propose a taxonomy for high-severity risks with a high level of granularity. Both taxonomies are valuable for identifying the full spectrum of systemic risks, as defined in Recital 110. Beyond taxonomies and typologies, a Preliminary Hazard Analysis conducted by Khlaaf et al. (2021), as reported in Table 4 of their paper, can serve as a foundation for analyzing risks of future systems.

Analysis

The analysis of systemic risks requires a combination of methods from both the risk management literature and the evaluations and red-teaming literature. On the evaluations and red-teaming end, Barrett et al. (2024) propose the BRACE (Benchmark and Red team AI Capability Evaluation) framework which, accounting for the tradeoff between cost-effectiveness and accuracy, proposes an evaluation procedure integrating benchmarks and red-teaming tests to provide a framework which is both applicable and precise. By leveraging correlations between easy to run benchmarks and highly labor intensive red-teaming evaluations, they use benchmarks as a cheap proxy of precise evaluations. Once a predefined threshold is hit on benchmarks, the more expensive red-teaming evaluations can be run.

To make evaluation more forward-looking, Phuong et al. (2024) from Google DeepMind make significant contributions that can be applied to GPAI systemic risks analysis:

1. A metric to quantify the distance between a model's current capabilities and dangerous thresholds. This continuous measure assesses the amount of expert guidance required for the model to successfully complete dangerous tasks.
2. Experts forecast elicitation as a useful indicator of when we might hit highly dangerous capabilities thresholds. The aggregate prediction is overall that, along all the axis (cybersecurity, persuasion, self-proliferation) some dangerous capabilities thresholds would be hit between 2025 and 2029.

3. A probabilistic approach to quantifying risk, estimating the likelihood of an AI system to successfully execute a dangerous task.

To facilitate interoperability and collaboration among various government entities conducting risk analyses, organizations may consider adopting and building upon the Inspect framework. This comprehensive evaluation framework, open-sourced by the UK AI Safety Institute, provides a standardized approach to evaluating AI systems.

A comprehensive range of techniques from standard risk management can be applied to GPAI systemic risks, as described in Koessler et al. (2023). These techniques include event and fault tree analysis, bow tie analysis, system-theoretic process analysis (STPA) and causal mapping. While we hope that the field of AI risk assessment will over time explore those methods, we think that in the near term, the most crucial one will be the Delphi technique that elicits experts' judgments and forecasts. It will be crucial, among others, to turn the risk tolerance into capabilities thresholds that can then be referred to when running risk analysis and measurements on AI systems.

Mitigation

The definition of measures to mitigate risks below acceptable levels relies heavily on an adequate and detailed definition of acceptable levels of risks. Once this foundation has been established, the range of mitigations for GPAI with systemic risks will have to cover deployment and containment measures, as named in Anthropic's "Responsible Scaling Policy" (Anthropic, 2023), along with a range of safety by design measures (e.g., safer architectures, pretraining with human feedback), safety engineering measures (e.g., dataset filtering, fine-tuning, input & output filters, decisions regarding external tools the model is given access to) or organizational controls (e.g., monitoring, KYC, training for deployers and users). For several of these mitigation categories, we'll suggest what we consider to be some of the most relevant publications on the topic. Beyond our targeted recommendations, to our knowledge, the most comprehensive document the EU GPAI Codes of Practice can draw upon for mitigation requirements is the "Emerging Processes for Frontier AI Safety" report from the UK AI Safety Institute.

For organizational controls, defining an Acceptable Use Policy and monitoring its adoption as done by OpenAI and Anthropic, is an important baseline. A key question that the Codes of Practice must consider is how long it takes for the monitoring to detect a malicious account. Given the possibility for malicious actors to create multiple accounts and switch between them when banned, reducing the detection time to less than a day might be crucial for mitigating high-severity risks.

As evidenced by the numerous LLM use cases for cyber offense experimented with by top-tier cybersecurity actors (Microsoft, 2024) high-severity misuse risks are likely to be the primary concern for GPAI models with systemic risks. To effectively reduce these risks, deployment measures should prioritize the elimination of

"jailbreak" type of failures, which is a key path to high-severity misuse. This might entail a range of mitigations such as "circuit breakers" (Zou et al., 2024), fine-tuning (Anil et al., 2024), or input/output filtering (UK AISI, 2023)

Containment measures such as infosecurity measures have been covered extensively by the recent RAND report (Nevo et al., 2024). This report provides a detailed analysis of potential attack vectors for stealing model weights and outlines the measures that can be implemented to achieve a desired level of security. We recommend that the Codes of Practice heavily relies on this resource when defining requirements related to cybersecurity and information security.

Ensuring International Interoperability

Beyond the content, the Codes of Practice will also have to be structured to facilitate interoperability with other frameworks. In this part, we'll review existing major frameworks and what design choices could be made in the EU GPAI Codes of Practice to facilitate interoperability.

In the standardization space, ISO/IEC 23894:2023 is the main risk management standard applied to AI. It provides a standard backbone with limited details and uses the risk management backbone that we discussed above, which is very standard. To ensure interoperability with this international risk management standard, we recommend that the Codes of Practice uses a variation of this outline. This approach allows for the inclusion of many of the obligations contained in Article 53 and 55 of the Act. This will be complemented by JTC21 writing EU standards that aim at, as the EU Joint Research Center (JRC) writes, complementing ISO/IEC 23894 by making it more detailed, more relevant to EU values and fundamental rights, and more focused on AI systems as opposed to organizations only, on which 23894 is focused. The risk management European Norm (EN) is also likely to follow a similar structuration as ISO/IEC 23894 and the classic risk management steps.

In the voluntary commitment space, the two most important tracks that have made significant progress are the G7 Hiroshima Process Codes of Conduct and the AI Safety Summits that led to the Seoul Frontier AI Safety Commitments.

The Hiroshima Process Codes of Conduct presents great similarity with the requirements of the AI Act and contains details of implementation that should be considered as part of the EU Codes of Practice, including:

- Under commitment 3, a mechanism of transparency report including important details such as "Details of the evaluations conducted for potential safety, security and societal risks" or "The results of red-teaming conducted to evaluate the model's/system's fitness for moving beyond the development stage".
- Under commitment 5, an emphasis on a range of best risk management practices such as establishing "policies, procedures, and training to ensure that staff are familiar with their duties and the organization's risk management practices".

- Under commitment 6, a commitment to have a security vulnerability management process.

The Seoul Frontier AI Safety Commitments are more narrowly scoped but also more detailed. Their language would be a very good source of inspiration for GPAI Codes of Practice, especially regarding the setting of acceptable levels of risks, defined under Outcome 1, particularly in points 1 through 4. Those points outline very specific good practices to assess and set adequate thresholds, along with how mitigation measures fit in that scheme. The Seoul Frontier AI Safety Commitments also pursue similar goals as the EU AI Act obligations for GPAI with systemic risks.

Finally, the NIST AI Risk Management Framework is a high-level framework for the risk management of AI systems. Although its content is similar to what has been discussed above, it uses a different backbone structure centered around four key steps: "Govern", "Map", "Measure" and "Manage". We recommend at the end of the EU Codes of Practice effort to create a table of interoperability as done in Barrett et al. (2023), to show how the Codes maps to the NIST framework. However, we recommend against directly using this framework for the Codes of Practice itself, due to its lesser interoperability than the ISO 23894 one discussed above. Based on this NIST framework, the UC Berkeley Center for Long-Term Cybersecurity (CLTC) has produced an extensive risk management framework for advanced GPAI systems. This work can serve as a valuable resource for the GPAI Codes of Practice, particularly in the areas of evaluations and addressing risks that are currently difficult to measure precisely.

On Systemic Risks

Recital 110 of the AI Act provides insights into what the AI Act considers systemic risks. For a number of the risks emphasized in this Recital, we will provide references.

As we suggested above, the two taxonomies we consider most relevant in covering the breadth of systemic risks are Weidinger et al. (2021) and Hendrycks et al. (2023).

On model fairness, we suggest that the Codes of Practice rely on the literature review Gallegos et al. (2023). It provides a comprehensive overview of existing metrics, datasets and mitigations that exist and can be used for large language models (LLMs).

On chemical, biological, radiological and nuclear risks, we suggest that the Codes of Practice draw upon the methodology of the latest OpenAI study (OpenAI, 2024) (while improving upon the statistics used to interpret it). Additionally, longer-form studies like RAND's Mouton et al. (2024) are promising tools that the EU AI Office could use to monitor risks at the frontier. Relevant design choices could also be adapted from Esvelt et al. (2023), with the addition of a control group that has access only to the internet.

Regarding risks arising from models self-replicating, Ngo et al., (2022) provide a useful breakdown of the problems among LLMs that could lead to such an outcome. The former head of the Superalignment team at OpenAI has also written a useful post (Leike, 2024) characterizing the ways a model could self-replicate and the capabilities that should be monitored to specifically avoid that outcome.

Concerning offensive cyber capabilities, the release of a number of cyber offense use cases by Microsoft's report (Microsoft, 2024) and papers demonstrating existing misuse of frontier AI systems are noteworthy. These include GPT-4 autonomously hacking weakly defended websites (Fang et al. 2024), or GPT-4 finding in a testing environment zero-day vulnerabilities (Fang et al. 2024).