Centrum für Europäische Politik
FREIBURG | BERLIN

cepAdhoc

No 25 | 2024 28 November 2024

Towards Robust AI Governance in Europe

Technical Recommendations for the General-Purpose AI Code of Practice

Anselm Küsters

Source: DALL-E via ChatGPT prompting

Artificial Intelligence is transforming industries and everyday life, prompting the EU to develop a General-Purpose AI Code of Practice to guide its safe and ethical use. This analysis examines the first draft of the Code, suggesting technical improvements to better support innovation, address systemic risks, and protect users.

- The Code's principles of alignment with EU values, proportionality, and future-proofing are laudable, but need to be better operationalised. In order to effectively support SMEs and foster innovation, the Code should provide concrete examples of simplified compliance pathways and take into account the specific challenges faced by open-source providers. In addition, the Code must explicitly prioritise freedom of expression and access to information, given that GPAI models are becoming fundamental platforms for internet interaction.
- The taxonomy of systemic risks should be expanded to include cascading and spillover effects, which are critical for comprehensive risk identification in interconnected AI systems. Interdisciplinary methodologies from complex systems analysis, network theory, and behavioural economics can enrich traditional risk assessment processes. The Code should also establish clear severity levels for risks and ensure high scientific rigour in assessments, possibly using metrics for inter-rater agreement.
- Continuous risk assessment during deployment is essential and should include systematic data collection on human-AI interactions, while respecting user privacy. The Code's emphasis on the use of "if-then" structures for risk mitigation measures and robust KPIs might enable automated self-assessments, but care must be taken to avoid oversimplification. Transparency requirements for downstream providers should be tiered according to their capabilities, while energy consumption reporting should be encouraged.

Content

1 Introduction: The General-Purpose AI Code of Practice 3
2 General observations 4
2.1 Alignment with Union principles and future-proofing 4
2.2 Proportionality and support for start-ups, SMEs, and open-source providers 5
2.3 GPAI as the new platform shift and freedom of expression 5
3 Working Group 2 observations 6
3.1 Taxonomy of systemic risks 6
3.2 Inclusion of „loss of control" 7
3.3 Dangerous model propensities and AI collusion 8
3.4 Safety and Security Framework (SSF) 8
3.5 Risk analysis methodology 9
3.6 Tiers of severity 9
3.7 Evidence collection and evaluation 10
3.8 Risk assessment lifecycle and user interaction data 10
4 Granular observations 11
4.1 Conditional risk mitigation measures using "if-then" structures 11
4.2 Robustness against misspecification 11
4.3 Transparency requirements for downstream providers 11
4.4 Detailed data transparency for developers 12
4.5 Energy consumption reporting 12
4.6 Limitations of robots.txt in data ownership 13
4.7 Safety mitigations through (mandated) system prompts 13
4.8 Innovative formats for Safety and Security Reports 14
4.9 Automating development and deployment decisions 15
4.10 Defining serious incidents 15
4.11 Sanctions for missed notifications 16
4.12 Dynamic classification criteria beyond compute thresholds 16
5 Conclusion 17

Figures

Fig. 1: Empirical example of using system prompts to improve model behaviour 14

1 Introduction: The General-Purpose AI Code of Practice

In November 2024, the European Union (EU)'s proactive approach to regulating Artificial Intelligence (AI) has reached another milestone with the publication of the first draft of the General-Purpose AI Code of Practice.[1] This draft, prepared by independent experts and facilitated by the European AI Office, aims to align with and support the forthcoming EU AI Act. The latter defines a General Purpose AI (GPAI) model as any AI model "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".[2] Such models, typically large language models (LLMs) like GPT-4 or image generation models like DALL-E, are thus characterized by their significant capabilities and ability to be integrated into diverse downstream applications. Here, the AI Act adopts a risk-based regulatory approach. While all GPAI providers face basic obligations, those deemed to pose "systemic risk" (presumed for models trained with over 10^25 floating point operations) are subject to additional requirements. As GPAI models become increasingly important in everyday life, they require a robust framework to ensure their safe use.

The current stakeholder consultation process on the above-mentioned "Code of Practice" aims to harmonise the efforts of developers, regulators, and users to address the novel risks posed by GPAI models while contributing to legal certainty and reducing bureaucratic burdens. Involving nearly 1,000 participants, the consultation aims to capture diverse perspectives and expertise, which is essential given the rapidly evolving nature of AI technologies. The Centre for European Policy (cep) is an active contributor to Working Group 2, focusing on the taxonomy of risks and the size and capacity of providers. This engagement is rooted in the belief that effective AI regulation must balance innovation with safeguards to foster an environment in which AI can thrive responsibly. Policies must reflect both the technological nuances of AI and the socio-economic realities of stakeholders. In particular, effectively regulating AI requires recognising its unique characteristics, namely exponential growth, autonomous iterative development, and often opaque internal processes ("black box" nature).[3] Traditional regulatory models may not be sufficient; instead, we advocate an iterative, bottom-up approach that promotes AI literacy and minimises systemic risks. Such an approach must be adaptive and future-proof.

This policy brief summarises our initial round of feedback on the first draft of the Code of Practice on General Purpose AI (the "Code"). Section 2 provides general observations on the overarching principles that shape the Code, emphasising its alignment with EU values, proportionality, and future-proofing. Section 3 discusses specific provisions related to Working Group 2, focusing on risk identification and assessment for systemic risks, and analyses the taxonomy of systemic risks and risk assessment methodologies. Section 4 presents more detailed observations that go beyond the focus of Working Group 2, addressing various "Measures" listed in the Code and suggesting improvements based on the publicly available draft. It is important to note that this contribution is limited to general observations on

---

1 The draft is accessible here: https://digital-strategy.ec.europa.eu/en/library/first-draft-general-purpose-ai-code-practice-published-written-independent-experts (in short: "Code of Conduct, 1st draft").

2 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, PE/24/2024/REV/1, OJ L, 2024/1689, 12.7.2024 (in short: "EU AI Act"), Art. 3, para. 63.

3 See, e.g., Küsters and Vöpel (2023), https://commongroundeurope.eu/blog/vorfahrtsregeln-stattstoppschilder-warum-europa-denanschluss-bei-kuenstlicherintelligenz-gerade-jetzt-nichtverlieren-darf/.

the first publicly available draft. As a member of the Working Group, the author, along with other academic members, has agreed to maintain the confidentiality of internal discussions during meetings.

2 General observations

2.1 Alignment with Union principles and future-proofing

The Code of Practice is commendably guided by six key considerations, which are set out at the beginning of the document:

- Alignment with the Union's principles and values
- Alignment with the AI Act and international approaches
- Proportionality to risks
- Proportionality to the size and capabilities of providers
- Support and growth of the AI safety ecosystem
- Future-proofing

From this list, the emphasis on future-proofing is particularly important. Given the exponential development and black-box nature of AI, regulatory frameworks need to be dynamic and adaptable.[4] The draft rightly suggests that "a balance should be struck between concrete requirements and flexibility to adapt and update rules as technology and industry evolve".[5] This approach is in line with academic perspectives on agile governance in AI, which advocate for regulatory mechanisms that can evolve with technological progress.[6] To operationalise future-proofing, the Code could feature, beyond the promised regular review cycles, mechanisms such as sunset clauses for specific rules or adaptive compliance pathways that take into account the emergence of new technologies. As noted in the draft Code itself, references to dynamic sources such as incident databases and consensus standards might help ensure that providers are kept abreast of the latest developments and best practices.

The first draft made available in mid-November acknowledges that it "does not yet include a section on how the Code will be reviewed and updated – this will be included in later iterations of the draft Code".[7] The intention is to develop a mechanism for ongoing review and updating to ensure that the Code remains relevant and effective. In this regard, a transparent and accessible review process is essential to the long-term success of the Code. Establishing a dedicated portal for civil society input, where stakeholders can easily submit feedback and share examples (such as screenshots of problems with GPAI interfaces like chatbots), would encourage inclusive participation. Frequent iterations are critical; waiting too long between updates could render the Code obsolete due to the rapid pace of AI development. Transparency in the review process enhances trust and accountability and thereby encourages active stakeholder engagement. This approach is consistent with adaptive governance models, which emphasise iterative policy-making in dynamic technological environments.[8]

---

4 Küsters (2024), https://www.cep.eu/eu-topics/details/anticipating-ai-instead-of-preventing-it.html.

5 Code of Conduct, 1st draft, p. 4.

6 For an overview, see: Araz Taeihagh, 'Governance of Artificial Intelligence', *Policy and Society* 40, no. 2 (3 April 2021): 137–57, https://doi.org/10.1080/14494035.2021.1928377.

7 Code of Conduct, 1st draft, p. 3.

8 For the case of autonomous vehicles, see: Araz Taeihagh and Hazel Si Min Lim, 'Governing Autonomous Vehicles: Emerging Responses for Safety, Liability, Privacy, Cybersecurity, and Industry Risks', *Transport Reviews* 39, no. 1 (2 January 2019): 103–28, https://doi.org/10.1080/01441647.2018.1494640.

2.2 Proportionality and support for start-ups, SMEs, and open-source providers

The draft appropriately recognises the diverse landscape of AI providers, emphasising that "obligations applicable to providers of general-purpose AI models should take due account of the size of the provider of general-purpose AI models and allow for simplified compliance pathways for SMEs and start-ups with fewer financial resources than those at the forefront of AI development."[9] This differentiation is indeed crucial to fostering innovation in the long run. However, to fully realise this goal, the Code should provide concrete examples of how measures can be simplified for SMEs. For example, offering standardised compliance templates, concrete exemptions from certain reporting obligations, subsidised access to third-party testing frameworks or facilities, and technical assistance programmes for updating AI models could reduce barriers to compliance. As young firms undertake riskier innovation,[10] tailored regulatory support can significantly improve their ability to innovate responsibly.

In a similar vein, the Code states that "Measures, sub-measures and KPIs [Key Performance Indicators] should be proportionate to the risks, meaning that they should be (a) appropriate to achieve the objective sought, (b) necessary to achieve the objective sought, and (c) not impose a burden that is excessive in relation to the objective sought".[11] This is a sound principle, recognising the need to balance risk reduction with the practical realities of implementation. Industry stakeholders are likely to appreciate this approach as it reduces unnecessary compliance costs and encourages innovation. However, the real impact will depend on how the principle is operationalised. Recognizing the "positive impact that open-source models have had on the development of the AI safety ecosystem,"[12] the draft wisely considers exemptions and tailored obligations for open-source providers. As experts have noted, open-source AI plays a pivotal role in democratizing access to AI technologies without increasing risks at the margin.[13] Still, the Code should delineate thresholds for when open-source models are subject to certain regulations, perhaps based on their deployment scale or potential for "severe" impact.

Overall, a key question going forward will be whether SMEs and open-source vendors should be supported by a) exempting them from some rules (e.g. burdensome transparency obligations) or b) merely facilitating their compliance with rules, which are the same for each GPAI model. From a competitiveness perspective, option a) might be preferable, while a strict focus on AI safety would support option b), as the GPAI models of smaller providers might still pose systematic risks. From the current draft, it appears that the AI Office is leaning more towards the second approach (option b).

2.3 GPAI as the new platform shift and freedom of expression

LLMs and other GPAI systems are rapidly emerging as fundamental platforms for information access and interaction on the internet, marking a significant paradigm shift in the way we consume and process information. While this shift is promising in many ways, it also raises critical concerns about preserving freedom of expression and equitable access to information. As Calvet-Bademunt argues, there is an urgent need to address the potential for AI systems to inadvertently censor or bias information, a concern that will become more pressing as these technologies become more integrated into our daily

[9] Code of Conduct, 1st draft, p. 4.

[10] Alex Coad, Agustí Segarra, and Mercedes Teruel, 'Innovation and Firm Growth: Does Firm Age Play a Role?', *Research Policy* 45, no. 2 (March 2016): 387–400, https://doi.org/10.1016/j.respol.2015.10.015.

[11] Code of Conduct, 1st draft, p. 3.

[12] Code of Conduct, 1st draft, p. 4.

[13] Sayash Kapoor et al., 'On the Societal Impact of Open Foundation Models' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2403.07918.

lives.[14] In this context, the development of the EU's GPAI Code of Practice is a critical opportunity to safeguard freedom of expression within the global framework of AI governance. As Calvet-Bademunt concludes, the Code should explicitly address the systemic risks associated with content moderation, information filtering, and the potential for AI to significantly influence public discourse. In particular, it must go beyond vague references to "fundamental rights" and explicitly prioritise freedom of expression. Drawing on human rights standards, the Code should require AI providers to develop clear, detailed, and publicly available use policies. In addition, any content restrictions implemented by AI systems should adhere to the principles of legitimacy and proportionality, ensuring that restrictions on speech are well justified and not overly broad. Technical risk mitigation strategies should prioritise methods such as prebunking, debunking, and counter-speech over outright censorship when addressing issues such as misinformation or biases. This approach would preserve the ability of researchers and the general public to explore a wide range of ideas using AI tools.

3 Working Group 2 observations

3.1 Taxonomy of systemic risks

Measure 6 introduces a taxonomy to identify and categorise systemic risks associated with general purpose AI models. It aims to provide a structured framework for vendors to identify potential risks that could have widespread impact. As of writing, this taxonomy lists risks related to: offensive cyber capabilities (such as vulnerability discovery or exploitation); dual-use science enabling chemical, biological, radiological, and nuclear weapons attacks through weapons development, design, acquisition, or use; issues stemming from the inability to control powerful autonomous AI models ("loss of control"); the potential for accelerated and unpredictable AI development via automated use of models for AI research and development; the facilitation of large-scale persuasion and manipulation, including disinformation or misinformation that threatens democratic values and human rights; and large-scale illegal discrimination against individuals, communities, or societies.[15]

However, the current taxonomy outlined in Measure 6.1 **does not explicitly include cascading or spill-over effects**, which are critical in times of rapid technological and societal change. These effects refer to situations where a failure or unintended consequence in one part of the system triggers a chain reaction that leads to broader systemic problems. Our previous cep research has shown that AI models trained on historical data may fail to perform reliably during unprecedented events in a "polycrisis", potentially affecting other critical parts of a system – a phenomenon observed during the COVID-19 pandemic, when predictive models failed to account for sudden shifts in consumer behaviour and market dynamics, with consequences ranging from automated credit scoring to the deployment of police forces.[16] The absence of cascading risks in the taxonomy may lead to an underestimation of the potential for GPAI models to contribute to systemic failures, particularly in interconnected systems such as finance, healthcare, or other critical infrastructure. This omission is surprising given that a subsequent section emphasises that, in addition to the inherent capabilities of AI models, the socio-technical contexts in which they operate should also be considered when assessing systemic risks (Measure 6.3.3).

---

[14] Here and in the remainder of this section I draw on: Safeguarding Freedom of Expression in the AI Era | TechPolicy.Press.

[15] Code of Conduct, 1st draft, p. 17.

[16] Anselm Küsters, 'AI as Systemic Risk in a Polycrisis', cepAdhoc (Berlin: Centre for European Policy, 13 December 2022), https://www.cep.eu/eu-topics/details/ai-as-systemic-risk-in-a-polycrisis-cepadhoc.html.

Such a holistic approach is supported by socio-technical systems theory, which advocates analysing technological systems within their broader social contexts.[17] The inclusion of cascading/spill-over effects in the taxonomy of systemic risks is also justified by complexity theory, which has recently been applied to AI safety.[18] As noted above, cascading effects represent a unique category of risk that can amplify and propagate across different sectors, potentially leading to unforeseen, far-reaching consequences. This is crucial in the context of complex AI systems that are increasingly interconnected (e.g. via APIs). In this sense, the Code should move away from a reductionist approach that seeks to understand AI systems and their risks by breaking them down into their component parts, and instead use other methods (see section 3.5 below) to analyse complex AI systems that consist of many independent parts that interact to produce effects that are greater than their sum. By incorporating cascading effects into the taxonomy, the Code would provide a more robust framework for risk assessment in the face of rapid technology deployment and global links.

In response to the question of what relevant considerations or criteria are to be taken into account when defining whether a risk is systemic (page 18 of the draft Code), the inclusion of cascading effects justifies the following criteria:

- Interconnectivity: The extent to which the AI system interacts with other systems.
- Complexity: The potential for small changes to have significant effects due to system dynamics or a non-linear relationship between cause and effect.
- Time sensitivity: The speed with which effects can propagate through systems.

In sum, cascading and spillover effects should be included in the taxonomy of systematic risks in order to address both direct and indirect risks of wide-spread AI deployment, redirecting the attention of GPAI developers and politicians towards more effective, society-wide mitigation strategies.

More generally, there are several other theoretical risks that could be added to the taxonomy. These include environmental risks from widespread AI adoption or, as our cep research has argued, societal risks from rising unemployment if AI models are substitutive rather than complementary in their labour market effects.[19] This raises the question of why this particular list of risks was chosen by the AI Office. A more structured and transparent way of creating the taxonomy of risks would therefore be welcome. One way of doing this would be to link the identified risks more closely to the language of the original AI Act.

3.2 Inclusion of „loss of control"

In Measure 6.1, the Code identifies "loss of control" as a systemic risk,[20] likely referring to scenarios where providers are unable to control powerful autonomous AI models.

While there is a theoretical potential for AI models to become uncontrollable, the framing of this risk requires careful consideration. The notion of AI systems leading to a "loss of control" often stems from speculative narratives of Artificial General Intelligence (AGI) surpassing human capabilities, as

[17] Gordon Baxter and Ian Sommerville, 'Socio-Technical Systems: From Design Methods to Systems Engineering', *Interacting with Computers* 23, no. 1 (January 2011): 4–17, https://doi.org/10.1016/j.intcom.2010.07.003.

[18] Dan Hendrycks (forthcoming), Introduction to AI Safety, Ethics, and Society, chapter 5, available here: 5.1: Complex Systems | AI Safety, Ethics, and Society Textbook.

[19] Küsters and Poli (2024), Resisting or Rebooting the Rise of the Robots? (cepStudie) | cep - Centrum für europäische Politik.

[20] Code of Conduct, 1st draft, p. 17.

discussed by Bostrom.[21] However, critical AI scholars like Whittaker have repeatedly noted that focusing on such speculative risks can divert attention from immediate and verifiable issues such as algorithmic bias, privacy violations, and security vulnerabilities.[22] A more balanced taxonomy would perhaps omit this issue in order to focus on explainability, which can mitigate risks associated with unintended behaviours without invoking speculative scenarios.[23]

3.3 Dangerous model propensities and AI collusion

Measure 6.3.2, which lists dangerous model propensities, explicitly addresses the potential for AI models to engage in collusion with other AI systems.[24]

The idea that AI models could "collude" has been explored in the context of competition law. Scholars such as Ezrachi and Stucke have theorised that AI could enable tacit collusion, where algorithms independently learn to coordinate pricing strategies without explicit communication.[25] While this conceptual argument has been prominent for a number of years, empirical evidence of AI-driven collusion has been limited. Subsequent studies suggest that while AI might facilitate collusion under certain conditions,[26] market dynamics and the diversity of algorithms used by different firms make widespread collusion unlikely. As noted by Dorner, "the literature is often lacking the perspective of computer scientists, and seems to regularly overestimate the applicability of recent progress in machine learning to the complex coordination problem firms face in forming cartels".[27] Therefore, while it is prudent to consider collusion as a potential risk, the Code should base its mandates on evidence-based assessments to avoid overemphasising speculative risks.

3.4 Safety and Security Framework (SSF)

Measure 7 emphasises that providers of general-purpose AI models should implement a comprehensive Safety and Security Framework (SSF). This framework is designed to assess and mitigate systemic risks associated with AI models, particularly those with high-impact capabilities. It recognises that less comprehensive measures may be sufficient where a new model has capabilities similar to those already safely deployed, without significant risks materialising.[28]

This last provision is particularly beneficial for researchers and SMEs, who are often playing catch-up with large, well-funded AI labs run by large technology companies. The widening gap between industry leaders and academic or smaller institutions has been documented repeatedly through metrics such as top-cited papers, patents, and contributions to leading AI conferences.[29] SMEs and academic institutions typically lack the extensive (computing and human) resources required for the most rigorous

---

[21] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Reprinted with corrections 2017 (Oxford, United Kingdom: Oxford University Press, 2017).

[22] See: https://thebulletin.org/2024/07/three-key-misconceptions-in-the-debate-about-ai-and-existential-risk/.

[23] Finale Doshi-Velez and Been Kim, 'Towards A Rigorous Science of Interpretable Machine Learning' (arXiv, 2017), https://doi.org/10.48550/ARXIV.1702.08608.

[24] Code of Conduct, 1st draft, p. 21.

[25] Ariel Ezrachi and Maurice E. Stucke, *Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy* (Cambridge, Massachusetts: Harvard University Press, 2016).

[26] Stephanie Assad et al., 'Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market', *SSRN Electronic Journal*, 2020, https://doi.org/10.2139/ssrn.3682021.

[27] Florian E. Dorner, 'Algorithmic Collusion: A Critical Review' (arXiv, 2021), https://doi.org/10.48550/ARXIV.2110.04740.

[28] Code of Conduct, 1st draft, p. 21.

[29] Nur Ahmed and Muntasir Wahed, 'The De-Democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research' (arXiv, 2020), https://doi.org/10.48550/ARXIV.2010.15581.

robustness checks performed by state-of-the-art AI developers. By allowing for proportionality in compliance measures, the Code recognises the resource constraints faced by these entities. However, it is important to monitor this approach over time as the industry landscape evolves and SMEs potentially develop more advanced models that may still warrant stricter oversight.

3.5 Risk analysis methodology

To effectively address the potential risks associated with AI systems, providers need to employ robust risk analysis methodologies (Sub-Measure 9.1). These approaches should not only identify potential pathways through which their AI models could create systemic risks, but also assess both the likelihood and impact of those risks.[30] Recent research in AI safety has begun to view AI as a complex system, exploring the unexpected behaviours such systems often exhibit.[31] This perspective challenges the so-called reductionist paradigm, which assumes that systems can be understood by breaking them down into their component parts, since AI systems consist of many independent parts that interact to produce effects greater than their sum. Consequently, the effects of AI on society, itself a complex system, are inherently difficult to predict, necessitating more sophisticated analytical approaches.

The traditional risk analysis techniques referenced by the draft Code, while valuable, may thus not capture the full complexity of AI systems and their interactions with human users and societal structures. It is therefore crucial to incorporate methodologies from other disciplines to enrich the assessment process. Complex systems analysis, as proposed by Holland for "complex adaptive systems", i.e. systems that involve many components that adapt or learn as they interact,[32] can provide insights into emergent behaviour in AI systems and help anticipate unforeseen consequences that may arise from the detailed interplay of system components. Likewise, network theory might provide tools for analysing how interconnected systems can propagate risk, which is particularly relevant in the "polycrisis" context mentioned earlier.[33] In addition, insights from behavioural economics, pioneered by researchers such as Tversky and Kahneman,[34] can help anticipate how users may interact with GPAI models in unintended ways due to cognitive biases, potentially exacerbating or creating entirely novel risks.

3.6 Tiers of severity

According to Measure 9.3, GPAI providers shall classify identified risks into levels of severity, with at least one level where risks are unacceptable without safeguards.[35] Creating such severity levels facilitates prioritisation in risk management.

While there is, at least to the author's knowledge, no universally accepted standard for these tiers in AI, providers can draw on systems thinking as well as established risk matrices used in safety-critical

[30] Code of Conduct, 1st draft, p. 22.
[31] Dan Hendrycks (2024), Introduction to AI Safety, Ethics, and Society, chapter 5, available online.
[32] John H. Holland, 'Studying Complex Adaptive Systems', *Journal of Systems Science and Complexity* 19, no. 1 (March 2006): 1–8, https://doi.org/10.1007/s11424-006-0001-z.
[33] Küsters, 'AI as Systemic Risk in a Polycrisis'.
[34] Amos Tversky and Daniel Kahneman, 'Judgment under Uncertainty: Heuristics and Biases: Biases in Judgments Reveal Some Heuristics of Thinking under Uncertainty.', *Science* 185, no. 4157 (27 September 1974): 1124–31, https://doi.org/10.1126/science.185.4157.1124.
[35] Code of Conduct, 1st draft, p. 22.

industries.[36] Developing a consensus on the concept of "severity" should go beyond the proposed approach of multiplying a specific risk with its probability and include qualitative criteria such as:

- Impact on human rights
- Scale, i.e. the number of people or systems potentially affected
- Irreversibility, i.e. whether the consequences can be undone

However, public concern, i.e. the level of societal anxiety associated with a particular risk, should not influence the definition of "severity", as public discourse is currently heavily influenced by dystopian science fiction narratives (see also section 3.2 above).

3.7 Evidence collection and evaluation

Measure 10 commits signatories to an ongoing process of gathering evidence on systemic risk, using a range of methods including forecasting, benchmarking, red-teaming, and simulation.[37]

Crucially, the choice of appropriate assessment method depends on the specific context. Certain risks are better assessed using specific methods. For example, security vulnerabilities may require adversarial testing, while bias may be assessed using fairness benchmarks.[38] However, computing power and expertise may limit the choice of methods, particularly for SMEs. To determine whether an evaluation is thorough, metrics analogous to Fleiss' Kappa could be used to assess inter-rater agreement in qualitative evaluations. Fleiss' Kappa is a statistical measure that allows for the assessment of agreement among multiple evaluators, which might be crucial when dealing with complex AI systems where subjective judgments could play a role in risk assessment. By quantifying the level of agreement between different evaluators, Fleiss' Kappa or similar measures provide a metric for the reliability of the evaluation process itself, adding an extra layer of rigor to the risk assessment methodology.

3.8 Risk assessment lifecycle and user interaction data

Measure 11 requires signatories to continuously assess risks throughout the lifecycle of the AI model, particularly during deployment. This includes updating risk assessments at least every six months or when significant changes occur, taking into account evidence from model monitoring.[39]

The emphasis on assessment during the deployment phase is critical because many risks arise from actual user interactions with GPAI models. Static assessments cannot capture the dynamic nature of how users engage with AI systems. A large body of research highlights that human-machine interactions can lead to unintended effects, including overreliance on AI suggestions or misinterpretation of results.[40] Understanding these dynamics is essential to mitigate risks in the long run. To effectively collect evidence on human-AI interactions, mode providers could implement channels for users to report issues such as misinterpretations or unintended model behaviours, analyse patterns in how users

[36] Nancy G. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety* (The MIT Press, 2012), https://doi.org/10.7551/mitpress/8179.001.0001.

[37] Code of Conduct, 1st draft, p. 23.

[38] Ninareh Mehrabi et al., 'A Survey on Bias and Fairness in Machine Learning', *ACM Computing Surveys* 54, no. 6 (31 July 2022): 1–35, https://doi.org/10.1145/3457607.

[39] Code of Conduct, 1st draft, p. 25.

[40] For an overview, which notes the role of AI's anthropomorphism, see: Ella Glikson and Anita Williams Woolley, 'Human Trust in Artificial Intelligence: Review of Empirical Research', *Academy of Management Annals* 14, no. 2 (July 2020): 627–60, https://doi.org/10.5465/annals.2018.0057.

interact with AI to identify common misunderstandings, or experiment with different model versions to assess user responses and model performance in real-world settings (A/B testing). However, as these examples illustrate, such a collection of user interaction data also raises privacy concerns. Compliance with the General Data Protection Regulation (GDPR) is mandatory. Anonymising data, obtaining informed consent, and ensuring data security are thus important prerequisites for this approach.

4 Granular observations

4.1 Conditional risk mitigation measures using "if-then" structures

The Code proposes tying "risk-mitigating Sub-Measures to risk-assessment KPIs, such as using 'if-then' requirements. For example, if a general-purpose AI model with systemic risk is assessed to have capability X, Y risk mitigations must be in place, guided by Z KPIs".[41]

This feature promises to improve the clarity of the envisaged compliance obligations and has some interesting implications from a digital policy perspective. By structuring requirements in an "if-then" format, providers can not only better understand and implement the necessary measures, but could even automate them, as this logic is similar to programming structures. Automating compliance checks could greatly benefit SMEs and developers by providing accessible tools to verify compliance with the Code without incurring high legal or expert costs. However, care must be taken to ensure that automated checks do not oversimplify the ethical considerations inherent in the use of GPAI models.

4.2 Robustness against misspecification

The Code emphasises that "Sub-Measures and KPIs should also be robust to circumvention or misspecification. The Code can accomplish this by, for example, avoiding the unnecessary use of proxy terms or metrics. The AI Office will monitor and review Sub-Measures and KPIs that may be susceptible to circumvention and other forms of misspecification".[42]

This is indeed a critical consideration as ambiguous or poorly defined metrics can lead, intentionally or unintentionally, to non-compliance or ethical lapses. Above all, ensuring robustness requires clear, unambiguous language and consideration of linguistic and cultural variations that may affect interpretation. International collaboration in AI development means that concepts and terminology may be understood differently in different regions. As Waber and co-authors have noted: "Because AI and related data regulations are rarely uniform across geographies, compliance can be difficult. To address this problem, companies need to develop a contextual global AI ethics model that prioritizes collaboration with local teams and stakeholders and devolves decision-making authority to those local teams."[43] Regular reviews by the AI Office can help identify and correct areas that are prone to misinterpretation.

4.3 Transparency requirements for downstream providers

Downstream providers integrating GPAI models are required to disclose technical details, including "a description of the model architecture, the type of model, the context size where appropriate, the total

[41] Code of Conduct, 1st draft, p. 3.
[42] Code of Conduct, 1st draft, p. 4.
[43] See: https://hbr.org/2024/08/how-companies-can-take-a-global-approach-to-ai-ethics.

number of model parameters and the number of parameters that are active during inference," as well as "versioned dependencies for required software and/or hardware".[44]

While transparency is essential (but not sufficient),[45] these requirements may be too demanding for downstream providers, especially SMEs or individual developers. Lately, the integration of GPAI models such as LLMs has been greatly simplified (think Meta's Llama series accessed via GitHub) and developers may not have in-depth knowledge of the underlying model architecture or parameters, especially when using pre-trained models or APIs. Imposing such detailed disclosure requirements could stifle innovation and exclude smaller players that lack the resources to comply. A tiered approach to transparency could be more effective, with the level of information required varying according to the concrete capabilities of the downstream providers.

4.4 Detailed data transparency for developers

Developers are required to disclose "specific information about the data used to train/test/validate the model, such as the fraction of the data that comes from different data sources, and the main characteristics of the training, testing and validation data".[46]

This feature of the current Code is highly commendable, as it promotes reproducibility and accountability in AI development. Transparent reporting of data sources and characteristics of training and testing is crucial for understanding potential biases and limitations of AI models.[47] There is growing concern in the academic community about the lack of reproducibility in machine learning research due to poor reporting.[48] In particular, the lack of separation between test data and training data has been criticised.[49] By requiring detailed data and training transparency, the Code aligns industry practices with academic standards, increasing the reliability of AI systems and fostering trust among users.

4.5 Energy consumption reporting

Developers must "detail which information and methodology they use to assess energy consumption […], in consistency with any delegated act adopted in accordance with Article 97 of the AI Act to detail measurement and calculation methodologies with a view to allowing for comparable and verifiable documentation".[50]

This requirement addresses environmental concerns associated with the computational demands of training large AI models. This is highly welcome, as our previous cep research has highlighted the significant carbon footprint of training cutting-edge GPAI models.[51] By requiring reporting of energy

[44] Code of Conduct, 1st draft, p. 11.

[45] Fort the inadequacy of transparency for understanding and governing algorithmic systems, see: Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability', *New Media & Society* 20, no. 3 (March 2018): 973–89, https://doi.org/10.1177/1461444816676645.

[46] Code of Conduct, 1st draft, p. 12.

[47] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ed. Sorelle A. Friedler and Christo Wilson, vol. 81, Proceedings of Machine Learning Research (PMLR, 2018), 77–91, https://proceedings.mlr.press/v81/buolamwini18a.html.

[48] Joelle Pineau et al., 'Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)', August 2021.

[49] Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart, *Text as Data: A New Framework for Machine Learning and the Social Sciences* (Princeton Oxford: Princeton University Press, 2022).

[50] Code of Conduct, 1st draft, p. 12.

[51] See: https://www.cep.eu/eu-topics/details/environment-takes-a-backseat-in-eu-digital-push.html.

consumption, the Code encourages developers to consider the environmental impact of their work, thereby providing a first step towards ensuring the EU's twin transition goals. Limiting this obligation to initial developers is appropriate, as they have control over the training process and access to relevant information. Downstream providers typically use pre-trained models and have thus less control over energy consumption.

4.6 Limitations of robots.txt in data ownership

The Code states that "[s]ignatories will only employ crawlers that read and follow instructions expressed in accordance with the Robot Exclusion Protocol (robots.txt)".[52]

While compliance with robots.txt is (at least in theory) standard practice, its effectiveness in the context of modern AI data collection is increasingly questionable. Robots.txt was developed in the early days of the internet and does not provide robust legal protection or technical enforcement mechanisms. Given the sophisticated methods used for data scraping today, relying solely on robots.txt may not be sufficient. As noted by Pierce: "For decades, robots.txt governed the behavior of web crawlers. But as unscrupulous AI companies seek out more and more data, the basic social contract of the web is falling apart."[53] The Code should consider mandating more advanced protocols or legal agreements to respect data ownership and privacy.

4.7 Safety mitigations through (mandated) system prompts

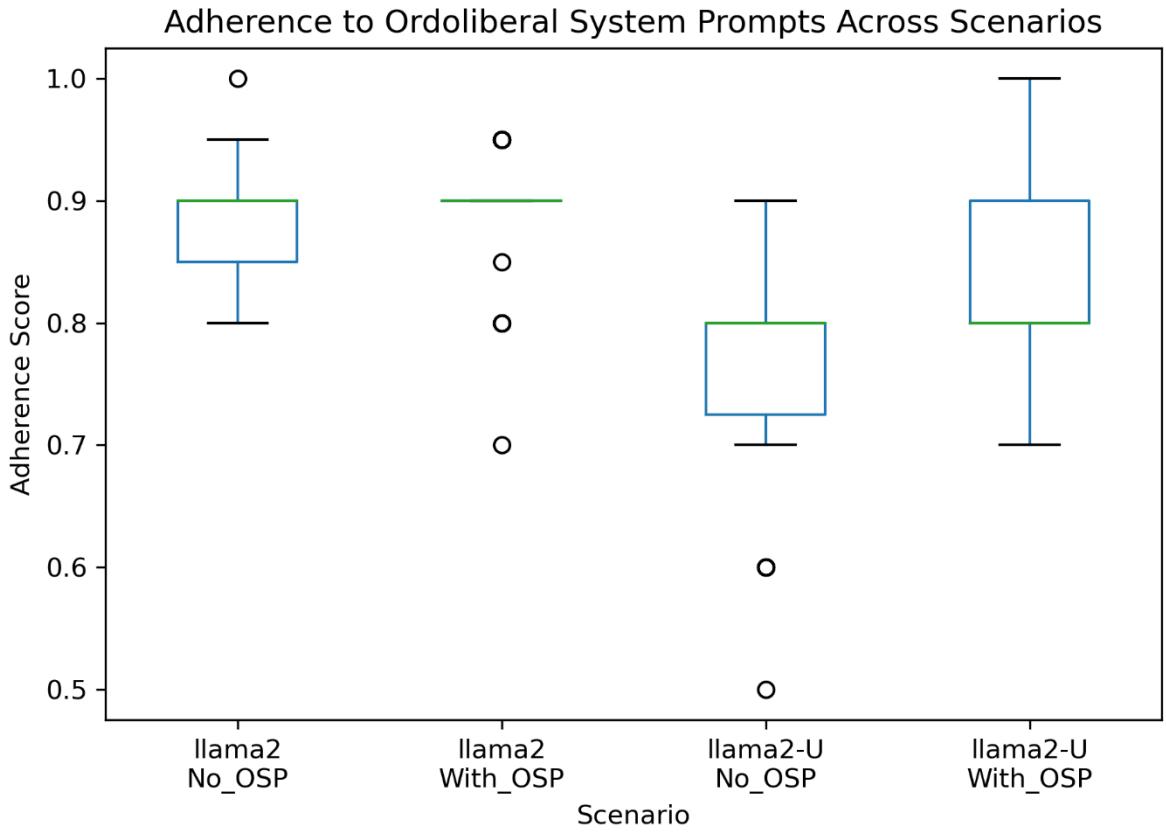
Measure 12 suggests that safety mitigations could include "behavioral modifications to a model", among other strategies.[54] But how to achieve this in practice, i.e. on a technical level?

As argued in our earlier cep research, the use of so-called "system prompts", also known as instruction tuning, might be a quick and largely effective way to guide AI models towards safe and intended behaviours.[55] For example, incorporating instructions that discourage the generation of harmful content or biases can help mitigate certain risks. By implication, mandating certain types of system prompts could partly standardise safety practices across the industry. However, overly rigid requirements could stifle creativity or lead to unintended consequences, such as limiting legitimate expressions of free speech (see section 2.3 above).

---

52 Code of Conduct, 1st draft, p. 15.
53 See: https://www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders.
54 Code of Conduct, 1st draft, p. 27.
55 See: https://www.cep.eu/eu-topics/details/in-search-of-laws-of-robotics.html.

Fig. 1: Empirical example of using system prompts to improve model behaviour



Source: Own figure based on unpublished research, forthcoming as Küsters/Wörsdörfer, "Exploring Laws of Robotics".

In a forthcoming paper entitled "Exploring Laws of Robotics: A Synthesis of Constitutional AI And Constitutional Economics", jointly written with Manuel Wörsdörfer, we investigated the use of system prompts, specifically self-designed ordoliberal system prompts (OSPs), as a practical method of guiding LLMs towards ethical and intended behaviour.[56] We conducted an empirical study comparing two versions of the Llama 2 model: the standard Llama 2 with built-in content moderation and Llama 2-Uncensored, which lacks internal moderation mechanisms. By posing ethical dilemma questions to both models, with and without the OSPs, we found that the OSPs had a negligible effect on the standard Llama 2, but significantly improved the ethical compliance of Llama 2-Uncensored. This suggests that while system prompts may not significantly alter models already equipped with security layers, they can effectively improve the behaviour of models without such mechanisms. Our results suggest that system prompts are an easily implementable and adaptable tool for improving the ethical behaviour of AI systems, particularly those that are uncensored or lack comprehensive adaptation strategies, although they are not a substitute for more robust pre-training adaptation methods.

4.8 Innovative formats for Safety and Security Reports

The Code also requires signatories to "create a Safety and Security Report (SSR) for any general-purpose AI model with systemic risk".[57]

[56] This paper is currently in the peer review process. An earlier version, without the empirical experiment but with the ordoliberal system prompts, is accessible here: https://www.cep.eu/eu-topics/details/in-search-of-laws-of-robotics.html.

[57] Code of Conduct, 1st draft, p. 28.

Such reports are typically limited to PDF documents, which are static and may not effectively convey complex, technical information. To avoid creating another overly detailed document that is difficult to understand, we suggest exploring other, more innovative formats. Exploring alternative formats such as computational reports using Jupyter notebooks could improve transparency and understanding. These formats allow for interactive exploration of code, data, and results, facilitating better scrutiny and reproducibility.[58] In particular, including elements such as code snippets, data visualisations, and executable cells could make SSRs more informative and accessible to stakeholders with different levels of expertise. This approach is consistent with open science practices and could improve the quality of documentation.[59] In conjunction with Measure 21 on documentation,[60] the Code could encourage the use of such formats to promote comprehensive and transparent reporting throughout the AI lifecycle.

4.9 Automating development and deployment decisions

Measure 14 states that "[s]ignatories commit to establish a process to decide whether to proceed or not with the development and deployment of a general-purpose AI model with systemic risk".[61]

Automating aspects of this decision-making process could improve objectivity and consistency. For example, pre-defined criteria or thresholds, as listed in other parts of the Code, could trigger automated actions, such as halting development if certain risk indicators exceed acceptable levels. This approach would help prevent biases or pressures that might lead developers to overlook potential risks, similar to so-called p-hacking in scientific research.[62] P-hacking refers to the manipulation of data analysis methods or selective reporting of results to artificially produce statistically significant results, often by exploiting researchers' degrees of freedom, which can lead to false positives and undermine the integrity of scientific research. Still, full automation may not be feasible due to the complexity of ethical considerations and the need for human judgement. A hybrid approach, where automation supports but does not replace human decision making, may be most effective. In conjunction with Measure 18 on serious incident reporting,[63] an integrated system could automatically flag incidents and ensure timely reporting to the EU AI Office and other relevant authorities.

4.10 Defining serious incidents

Measure 18 raises the question: "What does a serious incident entail?", and outlines obligations for tracking, documenting, and reporting such incidents.[64]

One sensible method for defining serious incidents could involve adopting a pre-registration protocol, similar to that used in clinical trials, where expected outcomes and risks are documented in advance, creating a division between "prediction" and "postdiction".[65] Deviations from expected outcomes

[58] Thomas Kluyver et al., 'Jupyter Notebooks ? A Publishing Format for Reproducible Computational Workflows', in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, ed. Fernando Loizides and Birgit Scmidt (IOS Press, 2016), 87–90, https://eprints.soton.ac.uk/403913/.

[59] Jeffrey M. Perkel, 'Why Jupyter Is Data Scientists' Computational Notebook of Choice', *Nature* 563, no. 7729 (November 2018): 145–46, https://doi.org/10.1038/d41586-018-07196-1.

[60] Code of Conduct, 1st draft, p. 34.

[61] Code of Conduct, 1st draft, p. 29.

[62] Megan L. Head et al., 'The Extent and Consequences of P-Hacking in Science', *PLOS Biology* 13, no. 3 (13 March 2015): e1002106, https://doi.org/10.1371/journal.pbio.1002106.

[63] Code of Conduct, 1st draft, p. 32.

[64] Code of Conduct, 1st draft, p. 32.

[65] Brian A. Nosek et al., 'The Preregistration Revolution', *Proceedings of the National Academy of Sciences* 115, no. 11 (13 March 2018): 2600–2606, https://doi.org/10.1073/pnas.1708274114.

could then be objectively identified as potential serious adverse events. Alternatively, incorporating ex-post expert judgement, e.g. through an ethics review board, could help determine the seriousness of incidents.[66]

4.11 Sanctions for missed notifications

Measure 20.1 allows the Commission to designate a model as having systemic risk if it becomes aware of such a model that was not notified.[67] However, there is no indication that there is any provision for concrete sanctions for failure to report.

Introducing sanctions for failure to notify relevant GPAI models to the AI Office could increase compliance, particularly among large AI providers that could have a significant impact on the market. Inspiration could come from EU competition law, where failure to notify mergers can result in significant fines, which act as a deterrent.[68] However, the imposition of sanctions must be carefully considered to avoid disproportionately affecting SMEs, open-source developers, or academic researchers who may lack resources or legal expertise. Exemption for these groups is warranted in order to recognise their different capabilities and to encourage innovation. Sanctions should focus on companies that have the capacity and responsibility to comply but choose not to. Awareness campaigns should help ensure that all providers understand their obligations.

4.12 Dynamic classification criteria beyond compute thresholds

The Code notes that "[t]he AI Office has the authority to update the classification criteria for determining whether a general-purpose model is presumed to have high-impact capabilities […]. How could it be written such that it is clear when providers should notify the AI Office of a model meeting new classification criteria?".[69]

Relying solely on computational thresholds for classification (e.g. models trained with over 10^25 floating point operations) is increasingly inadequate, as scholars have recently noted.[70] Advances in algorithmic efficiency and decentralised training methods mean that powerful models might be developed without exceeding high computational thresholds. Alternative criteria could include performance benchmarks on standardised tests or skill assessments. For example, referencing ever-evolving leaderboards such as the "Chatbot Arena", or benchmarks from the AI Alignment community,[71] can provide more dynamic, contextual assessments. Clear communication from the AI Office about updates to the classification criteria is essential.

---

[66] See also: Luciano Floridi et al., 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations', *Minds and Machines* 28, no. 4 (December 2018): 689–707, https://doi.org/10.1007/s11023-018-9482-5.

[67] Code of Conduct, 1st draft, p. 33.

[68] Ariel Ezrachi, *EU Competition Law: An Analytical Guide to the Leading Cases*, 5th ed. (Oxford: Hart, 2016).

[69] Code of Conduct, 1st draft, p. 34.

[70] Sara Hooker, 'On the Limitations of Compute Thresholds as a Governance Strategy' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.05694.

[71] Dan Hendrycks et al., 'Aligning AI With Shared Human Values' (arXiv, 2020), https://doi.org/10.48550/ARXIV.2008.02275.

5 Conclusion

The General-Purpose AI Code of Practice will help the EU steer AI model developers towards fostering a safe, trustworthy, and innovative AI ecosystem in Europe. This analysis, based on the first publicly available draft, underscores the Codes strengths but also highlights areas where improvements could enhance its effectiveness. First, by embedding flexibility to adapt to technological advances, the Code recognises the rapid evolution of AI and the need for dynamic regulation. The principle of proportionality ensures that obligations are balanced against the risks and capabilities of providers, in particular supporting SMEs and start-ups. Moreover, explicitly recognising the contributions of open-source providers strengthens AI development in the long run and promotes inclusivity.

Second, the specific provisions related to Working Group 2 emphasise the importance of a comprehensive approach to systemic risk. Expanding the taxonomy to include cascading and spillover effects would be crucial to capture complex risk scenarios, especially in interconnected systems during a "polycrisis" scenario. Incorporating interdisciplinary methodologies enriches risk analysis by providing a multi-dimensional, more holistic understanding of potential threats. Establishing clear levels of severity helps to prioritise mitigation efforts, while ensuring high scientific rigour in assessments maintains trust and accountability. Continuous risk assessment throughout the AI lifecycle, particularly during deployment, addresses emerging risks from real-world use and human-AI interactions.

Finally, this feedback document outlines some practical recommendations for refining the Code. Establishing a transparent review and update process, with accessible pathways for civil society input, could ensure that the Code remains up-to-date and reflects diverse perspectives. Better balancing transparency requirements for downstream providers prevents undue burdens on SMEs and encourages broader participation in AI innovation. Incorporating innovative formats for safety and security reports, such as Jupyter notebooks, would improve clarity and facilitate deeper understanding among stakeholders. Automating aspects of the decision-making process for development and deployment could improve objectivity and avoid bias. Clear definitions of serious incidents and proportionate sanctions for non-compliance are crucial for a robust regulatory framework.

In conclusion, the first draft provides a solid foundation for the General-Purpose AI Code of Practice. The Code will be finalised by April 2025, nine months after the AI Act's entry into force on 1 August 2024. By addressing the highlighted areas for improvement, the Code could more effectively balance the promotion of innovation with the imperative of safety. This will foster a European AI landscape that not only drives technological progress, but also protects societal values and mitigates systemic risks. The Centre for European Policy (cep) will continue to contribute its expertise, ensuring that it serves the best interests of society.

Centrum für Europäische Politik
FREIBURG | BERLIN

Author:

Dr. Anselm Küsters, LL.M., Head of Division Digitalisation and New Technologies
kuesters@cep.eu

Centrum für Europäische Politik FREIBURG | BERLIN
Kaiser-Joseph-Straße 266 | D-79098 Freiburg
Schiffbauerdamm 40 Räume 4205/06 | D-10117 Berlin
Tel. + 49 761 38693-0

The **Centrum für Europäische Politik** FREIBURG | BERLIN, the **Centre de Politique Européenne** PARIS, and the **Centro Politiche Europee** ROMA form the **Centres for European Policy Network** FREIBURG | BERLIN | PARIS | ROMA.