CDT Europe is a Brussels-based civil society organisation with international presence that works to uphold human rights in tech policy through research and advocacy. We advocate for the adoption of robust, technically-informed solutions for the effective regulation and governance of AI systems and models to ensure their compatibility with fundamental rights and to advance the interests of communities most impacted by AI.

We commend the Commission, the AI Office and the Chairs for their efforts in drafting the Code of Practice, and are grateful for the opportunity to provide the following feedback. We hope that our submission will assist the drafters in operationalising key fundamental rights protections and transparency principles.

We offer our written comments in two separate documents.

1. **The draft taxonomy of systemic risks is inadequately focused on evidenced current risks to fundamental rights in favor of theoretical safety risks of unclear likelihood.**

CDT appreciates that the draft Code of Practice does not aim to provide an exhaustive list of systemic risks, and this non-exclusive intent should be preserved in subsequent drafts. It is also appropriate that the Code consider risks based on both their likelihood of occurring and their likely severity if they do occur, as per the definition of risk in Article 3(2) of the AI Act. However, it is important to recognise that providers are most likely to–indeed, are required to–prioritise the risks specifically highlighted by the Code in its non-exclusive taxonomy of systemic risks. As a result, we are concerned that the draft list of systemic risks is overly focused on potentially severe risks where the likelihood of that risk occurring–or even the definition of that risk–are unclear due to lack of meaningful evidence specific to that risk. It should instead prioritize more immediate and evidenced risks to fundamental rights.

In particular, the list seems to heavily favor what may be characterised as national security, "catastrophic," "emergent," or "existential" risks. For example, at the top of the list are cyber offence and chemical, biological, radiological and nuclear [CBRN] risks. However, there's a lack of clear evidence that current or imminent models are likely to pose a significant marginal risk in these areas compared to other existing technologies,[1] in which case spending significant resources on those risks from GPAIs at the expense of other more concrete risks would be a poor allocation of providers' limited capacities.

Similarly, broad and vague risks such as "loss of control" of "powerful autonomous [GPAIs]" or "loss of trust in media" due to "large-scale [GPAI-enabled] disinformation or misinformation"–although certainly the subject of much academic speculation–are still quite fuzzily conceptualised and lacking in clear evidence of severe risk.[2] That raises significant questions not only about how to evaluate or mitigate those risks but also about whether doing so would reduce the resources available to address more obvious and

[1] Sayash Kapoor et al., "On the Societal Impact of Open Foundation Models," *arXiv*, February 27, 2024, https://doi.org/10.48550/arXiv.2403.07918 (surveying current dearth of evidence that frontier models raise a substantial marginal risk in areas of mis/disinfo, biosecurity, cybersecurity, etc.); See also U.S. National Telecommunications and Information Administration, "Report on Dual-Use Foundation Models with Widely Available Model Weights," July 30, 2024, https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf (same).
[2] Ibid.; see also Kevin Bankston, "Comments to NTIA on Open Foundation Models," *Center for Democracy & Technology*, https://cdt.org/insights/cdt-comments-to-ntia-on-open-foundation-models/ esp. at pp. 22-24 (discussing "emergent" risks).

evidenced systemic risks–particularly in the case of SMEs and lesser-capitalised collaborative open source AI projects with relatively limited resources.

In stark contrast, issues of bias and discrimination are only addressed last in the taxonomy, and quite narrowly, despite a wealth of research highlighting current biases in existing general purpose models.[3] Meanwhile, other severe risks from, e.g., invasions of privacy or sexually abusive content, are left unaddressed despite clear evidence that general purpose models can and do pose systemic risks in these areas as well.[4]

This prioritization is backwards, treating speculative risks as a higher priority than actual risks being posed here and now, based on uncertain predictions about those theoretical risks' potentially catastrophic scope. Therefore we strongly recommend that the drafters reconsider how they are implicitly calculating the probability and severity of these threats, and make the following changes and additions to the taxonomy based on a more explicit, evidence-based weighing of risk probability and risk severity.

a) Expand the systemic risk related to discrimination

Particularly considering the evidence of serious and potentially intractable bias in current GPAI architectures, the likelihood that GPAIs will be integrated into a wide range of decision-making contexts, and the fundamental nature of the right against discrimination, the description of the systemic risk of discrimination should be broadened.

First, the "large-scale" qualifier is unnecessary considering the Act's recognition that one of the main reasons GPAIs pose outsized risk is because they are likely to be widely deployed in a broad range of contexts. It also risks the possibility that discrimination against smaller minority populations may be deprioritised, even though protecting minority populations is the core purpose of the right against discrimination.

Similarly, the restriction to "illegal" discrimination is unnecessarily narrow. First, the application of anti-discrimination law in this new technical context is unclear, and that lack of clarity may lead to providers pursuing unnecessarily narrow evaluations and mitigations. Second, equality bodies in the EU have long-recognised the limits of the protected characteristics recognised in EU law,[5] exploring additional grounds such as socio-economic status and health status.

---

[3] See, e.g., Valentin Hofmann, et al., "Dialect Prejudice Predicts AI Decisions About People's Character, Employability, and Criminality," *arXiv*, March 1, 2024, https://arxiv.org/abs/2403.00742; Jesutofunmi A. Omiye, et al., "Large Language Models Propagate Race-based Medicine," Npj Digital Medicine 6, no. 1, October 20, 2023, https://doi.org/10.1038/s41746-023-00939-z; and Lena Armstrong, et al., "The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring," arXiv, last revised May 7, 2024, https://arxiv.org/abs/2405.04412.

[4] See, e.g., Nicholas Carlini, et al, "Extracting Training Data from Diffusion Models," *arXiv*, January 30, 2023, https://doi.org/10.48550/arXiv.2301.13188 (demonstrating privacy risk of models regurgitating training data, which may include personal data); see also further discussion of sexually abusive content in following section.

[5] Equinet, "Expanding the List of Protected Grounds Within Anti-Discrimination Law in the EU", 2021, https://equineteurope.org/wp-content/uploads/2022/03/Expanding-the-List-of-Grounds-in-Non-discrimination-Law_Equinet-Report.pdf

Indeed, the AI Act itself favours an expansive interpretation of protected characteristics: the prohibited practice set out in Article 5(1)(b) acknowledges that socio-economic circumstances may lead to vulnerabilities, endorsing a holistic approach to discrimination that is not limited by currently recognised protected characteristics. Other systemic risks in the taxonomy do not hinge on illegality–indeed, as discussed in the next section, they often reach conduct and capabilities that may themselves be protected by fundamental rights–and there is no unique justification to limit the discrimination risk in this way.

A more appropriately broad risk category than "large-scale illegal discrimination" would address bias and discrimination in any highly consequential decision-making context, akin to the broad category of automated decisions that are encompassed in Article 20 of the General Data Protection Regulation, i.e., those that "produce[] legal effects concerning [a person] or similarly significantly affect[] him or her."

b) Include the generation of child sexual abuse material (CSAM) and nonconsensual intimate imagery (NCII) as systemic risks

The Code of Practice should expand the current taxonomy of systemic risks to include CSAM and NCII, which are widely known outcomes of the misuse of AI models. The relationship between AI models and CSAM is a complex one. Not only can CSAM be found in AI training datasets, but models trained on those datasets can generate CSAM, too.[6] The use of AI to generate CSAM has rightly become a growing area of concern,[7] leading researchers to raise the alarm and call for dedicated mitigation measures.[8] Careful consideration of the potential for an AI model to generate CSAM and necessary mitigations are crucial not least because current efforts to curtail CSAM documented to be effective have so far focussed on the identification of known CSAM,[9] and have had limited impacts in stemming AI-generated CSAM, which has continued to proliferate[10] and threatens to overwhelm child protection authorities.[11]

Similarly, the potential for AI models to be used to generate NCII, a type of image-based sexual abuse (IBSA), must be recognised in the Code of Practice. Research pre-dating the explosion of AI-generated NCII documented the serious mental and physical impacts of IBSA on its victims, noting the deeply gendered nature of its harms with victims being mostly women and girls.[12] The Digital Services Act, applicable to very large online platforms and search engines – where NCII is often disseminated – explicitly identifies gender-based violence as a category of systemic risk. It stands to reason that the AI Act's Code of Practice, set to govern the AI models which often *generate* this content, should do the same. This is consistent with the direction taken by industry: many AI developers have voluntarily

[6] David Thiel, "Investigation Finds AI Image Generation Models Trained on Child Abuse", December 20, 2023, https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse
[7] Internet Watch Foundation, "How AI is being abused to create CSAM", October 2023, https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf
[8] David Thiel, et al., "Generative ML and CSAM: Implications and Mitigations", Stanford Digital Repository, June 24, 2023, https://purl.stanford.edu/jv206yg3793.
[9] Google Safety Center, "NCMEC, Google and Image Hashing Technology", https://safety.google/intl/en_be/stories/hash-matching-to-help-ncmec/
[10] Internet Watch Foundation, "What has changed in the AI CSAM landscape?", July 2024, https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/
[11] Katie McQue, "AI is overpowering efforts to catch child predators, experts warn", *The Guardian*, July 18, 2024, https://www.theguardian.com/technology/article/2024/jul/18/ai-generated-images-child-predators
[12] Claire McGlynn and Erica Rackley, "Image-Based Sexual Abuse: More than Just 'Revenge Porn'", https://www.birmingham.ac.uk/Documents/college-artslaw/law/research/bham-law-spotlight-IBSA.pdf

committed to the newly-created IBSA principles, which require companies to invest in preventive action and be transparent about measures taken to address the creation and distribution of IBSA.[13]

c) Include invasions of privacy and mass surveillance as systemic risks

Models are trained on vast amounts of data - including personal data - often processed in the absence of a valid legal basis,[14] which raises important questions about the legality of data processing activities carried out by model providers which remain unanswered. For example, many AI models are trained on data scraped from the web, which poses serious challenges from a data protection perspective as scraping tends to collect data indiscriminately - including sensitive data - and the data subject is not aware that this processing is taking place. The processing of personal data, combined with the absence of individual notification, has led some regulators to find violations of GDPR in relation to the use of publicly scraped material .[15] In addition to this initial processing at the training stage, AI models are known to memorise and regurgitate personal data. Studies have shown that memorisation of a model grows significantly as the capacity of the model increases.[16] Given the inherent propensity of models categorised as posing a systemic risk under the AI Act to process massive amounts of data at multiple stages of model training, development and deployment, the list of systemic risks should explicitly include privacy infringements. Mass surveillance is intricately connected to the privacy issue: models present facts, including personal data, learned from patterns from multiple sources, and are therefore capable of processing personal data indiscriminately at a large scale.

d) Re-order the taxonomy from least to most speculative, or make clear the ordering is random.

Without clearer guidance, providers may read the ordering of systemic risks in the taxonomy to indicate some level of prioritisation. Either the drafters should clearly disclaim any such intent, or re-order to rank priority based on what risks are least speculative and most supported by specific evidence. A clearer prioritisation based on actual evidence of risk would help ensure that the systemic risks that are most likely to occur are also those most likely to be effectively managed, even when SME and open source GPAI providers are working with limited resources.

[13] IBSA Principles, https://ibsaprinciples.org/; see also Centre for Democracy and Technology, "Companies, Civil Society, Academics Announce Voluntary Principles to Combat Image-Based Sexual Abuse", September 12, 2024, https://cdt.org/press/companies-civil-society-academics-announce-voluntary-principles-to-combat-image-based-sexual-abuse/

[14] While model providers often rely on legitimate interest being an adequate legal basis, this remains an open question. See our reasoning in our submission to the French DPA: https://cdt.org/wp-content/uploads/2024/10/CDT-Europe-submission-to-CNIL_web-version.pdf

[15] Autoriteit Persoonsgegevens, Decision dated 16 May 2024, para. 117. Available at: https://www.autoriteitpersoonsgegevens.nl/en/current/dutch-dpa-imposes-a-fine-on-clearview-because-ofil legal-data-collection-for-facial-recognition

[16] Nicholas Carlini et al., "Quantifying memorization across neural language models", *arXiv*, last revised March 6, 2023, https://arxiv.org/abs/2202.07646

2. **The Code must more clearly highlight the importance of preserving the fundamental right to free expression, especially considering the problematic breadth and vagueness of some of the systemic risks identified in the taxonomy.**

The draft Code of Practice correctly calls for the protection of fundamental rights, which includes the right to freedom of expression, in the implementation of its requirements. However, the draft does not specifically mention free expression as a value to be preserved and protected. This is a notable omission because the line between free expression and activities giving rise to systemic risk is not always clear. Steps to mitigate many of the current systemic risks outlined within the Code could threaten free expression values in the design and use of general purpose AI tools if they are not properly scoped, and overly focusing on those risks at the expense of free expression could undermine the value of the use of GPAIs in many settings. Therefore we urge the drafters to amend the draft Code to more explicitly call for the protection of free expression in its overall implementation, and to more consciously consider the risks to free expression inherent in overly broad definitions of systemic risk.

As currently drafted, some of the systemic risks presented in the taxonomy are too broadly defined and could lead to unintended consequences for the free exchange of information, the pursuit of knowledge, and free expression generally. For example, chemical, biological, radiological, and nuclear risks intersect with the pursuit of scientific knowledge and advancement generally. The difference between analysis that leads to a lifesaving vaccine and analysis that instructs in the creation of a biological weapon is not always obvious. Similarly, the use of technology to persuade an audience does not always pose a significant risk of harm. Every political campaign or advertising campaign is an attempt to persuade. Similarly, the question of what constitutes harmful “disinformation” and “misinformation” is a highly contingent and uncertain one, and the answer often differs depending on one’s political ideology, making it difficult to identify whether certain content raises systemic risks or is a legitimate exercise of free expression–and raising the possibility that such legitimate expression could be over-censored in the hunt for untruths.

Potential additional systemic risks contemplated by the draft (and called for by CDT) could also implicate free expression. For example, while CSAM and NCII generation should be considered a systemic risk, policing that risk without adequate consideration of the consequences for free expression could impact, e.g., beneficial uses for sex education or even undermine the quality of the performance of the model when generating non-sexual content.

For these reasons, the definitions of systemic risks should be scoped more narrowly and clearly to explicitly account for the protection of free expression. Furthermore, the Code should explicitly highlight that decisions about systemic risk *mitigations* should also account for and seek to protect free expression. The draft rightly recognizes that the systemic risks it identifies could also include beneficial uses, and we encourage that the final Code state clearly that considerations of beneficial uses include the protection of human rights, particularly free expression, in the development and deployment of GPAI models. Considering the difficulty in identifying mis- and dis-information and drawing the line between beneficial scientific research and analysis and more harmful activities, the Code should encourage signatories to focus as much as possible on mitigations least likely to lead to over-censorship of legitimate expression, e.g. using behavioral analysis and other interventions that do not rely on content analysis. All mitigation decisions should consider impacts on the performance of the model and impacts on free expression and should be scoped to protect beneficial, rights-respecting uses as much as possible.

3. The required documentation in the draft Code lacks critical information for downstream providers seeking to deploy responsibly and should be expanded to address this gap.

The draft Code in Part II specifies the technical documentation that developers of general-purpose AI models must supply to both the AI Office and downstream providers. These requirements collectively include much of the information necessary for the AI Office and national competent authorities as well as downstream providers to understand the models' capabilities and limitations and fulfill their obligations. However, a key issue in the current draft is that many critical pieces of information—notably, details about the design specifications and training process, data validation and suitability assessment, and testing procedures—are required to be disclosed only to the AI Office. This inappropriately limits downstream deployers' ability to assess and effectively mitigate risks associated with these models.

a. The Code should require model providers to share documentation with deployers detailing the "design specification and training process."

The current draft of the Code does not require GPAI providers to share with downstream providers the documentation required by Annex XI sec. 1.2(b) on the design and training process. However, general-purpose model providers' decisions about design specification and training, such as the choice of objectives to optimise, play an important role in determining a model's suitability for specific use cases and significantly influence the model's behavior. Without insight into these design choices and general-purpose model providers' rationale for them, downstream deployers may lack the requisite understanding to evaluate whether the model will perform consistently with their applications' intended purposes.[1] Such an understanding may also be instrumental in helping deployers identify scenarios where the model may fall short in meeting their needs, either in failing to attain the necessary level of performance or in conveying unacceptable levels of risk. Any information about design specifications and training processes, therefore, should also be provided to downstream providers.

b. The Code should require that all "information on data used for training, testing, and validation"–including methods to detect unsuitable or biased data–be supplied to downstream providers.

While the current draft does require GPAIs to provide some documentation on aspects of training, testing, and validation data (as per Annex XI sec 1.2(c) and Annex XII 2(c)) to both the AI Office and downstream deployers, it only requires disclosure to the AI Office of "the methods used to detect unsuitability of data sources and any biases in the data" used for training, testing and validation. This information should be provided to downstream providers as well, as risks arising from unsuitable or biased data can cascade into downstream applications if not properly addressed.[2]

---

[1] Amy Winecoff and Miranda Bogen, "Report - Improving Governance Outcomes Through AI Documentation: Bridging Theory and Practice", *Centre for Democracy and Technology*, September 25, 2024, https://cdt.org/insights/report-improving-governance-outcomes-through-ai-documentation-bridging-theory-and-practice/

[2] Angelina Wang and Olga Russakovsky, "Overwriting Pretrained Bias with Finetuning Data", *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* 2023, https://openaccess.thecvf.com/content/ICCV2023/html/Wang_Overwriting_Pretrained_Bias_with_Finetuning_Data_ICCV_2023_paper.html

c. The Code should require that details on the "testing process and results thereof" be provided to downstream providers as well as the AI Office.

The draft Code fails to require GPAI providers to share with downstream deployers their documentation of testing procedures and results as per Article 53(1)(A). However, downstream deployers need this information about upstream evaluations, including "a description of the tests performed and the results of these tests," to determine whether these models meet their specific requirements. When no evaluations have been conducted by upstream model providers, it becomes even more critical for downstream providers to have this information, as the suitability of unevaluated models will be largely unknown to them.

d. The Code should require providers to document steps taken to address issues with datasets or evaluations of mitigation efforts

In addition, the draft Code does not impose any documentation requirements for general-purpose model providers on the actions they have taken to address identified issues with their datasets or evaluations of the effectiveness of these mitigation efforts. The updated Code should require general-purpose model developers to share with the AI Office and downstream providers documentation on their methods for identifying potential data-related issues, the mitigation measures implemented to address these issues, and the results of evaluations assessing the success of these measures in resolving the underlying problems.

e. The Code should require providers to furnish detailed documentation on their evaluation methods

Moreover, evaluation results alone offer limited value to downstream providers without additional granularity and context.[3] To fully understand the rigor and relevance of general-purpose model evaluations, general-purpose model providers should also supply both the AI Office and downstream providers with detailed documentation about their evaluation methods. This should include the general-purpose model provider's rationale behind selecting specific model properties and behaviors for evaluation, justifications for the chosen tests, and any known limitations or threats to the validity or reliability of their chosen tests. Providing documentation on subjective decisions made during development can render visible the normative assumptions underlying the model's development, which are often responsible for downstream harms.[4] Documentation on evaluations for deployers should also include guidance on interpreting the results in the context of downstream applications. Since general-purpose model evaluations are inherently disconnected from specific deployment contexts, this guidance can help bridge the gap and ensure downstream deployers can apply the findings effectively and correctly gauge sources of uncertainty about the provider's evaluations.[5]

---

[3] Ryan Burnell et al., "Rethink reporting of evaluation results in AI", *Science* 380 (6641), https://par.nsf.gov/servlets/purl/10448744

[4] Mireia Yurrita et al., "Generating Process-Centric Explanations to Enable Contestability in Algorithmic Decision-Making: Challenges and Opportunities", *arXiv*, May 1, 2023, https://arxiv.org/abs/2305.00739

[5] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems", *arXiv*, October 31, 2023, https://arxiv.org/pdf/2310.11986

Finally, and similar to data mitigations, if a general-purpose model provider has taken action based on evaluation results to address potential issues, they should explain their rationale for the chosen mitigation approach, describe how it was implemented, and report on the effectiveness of these efforts in resolving the issue. If no mitigation efforts were undertaken, it is equally important for general-purpose model providers to disclose this information to downstream providers since they will then need to address the issues entirely within their own implementations.

It appears possible that the drafters intended such information to make its way to downstream developers by virtue of being included in the Safety and Security Reports (SSRs) for the relevant models, which Measure 22 (Public Transparency) indicates should be published. However, the interaction of this requirement with Part II's documentation requirements is unclear, and the Code should clearly require that such information be included in the documentation required to be shared downstream.

Incorporating these additional documentation requirements into the next draft of the Code will ensure that downstream deployers have sufficient information about the evaluation and mitigation of limitations and risks of general-purpose models. This will better enable downstream providers to make more informed decisions about model selection and use, integration within their applications, and additional risk mitigations and evaluations they need to employ to address unresolved issues with their selected general-purpose model.

4. The Code should be clarified to more adequately address the special needs of open source releases.

We appreciate the draft's attempts, as required by the AI Act itself, to make special accommodations for SMEs and open source AI providers proportional to their oft-times limited resources compared to larger corporate or closed-model providers. We agree that, in our shared efforts to manage risks from highly capable models, we should seek to avoid unduly burdening smaller competitors or open scientific research efforts that could help foster more innovation and less centralisation of GPAI capacity amongst a handful of large companies. In furtherance of that goal, we highlight a few areas where greater clarity in the draft would be helpful.

a. As open source licenses typically do not restrict end uses, the Code should clarify that while transparency around any applicable acceptable use policy (AUP) is a requirement for open source providers, actually having one is not

The AI Act appropriately requires that the technical information to be provided to the AI Office and downstream providers include "the acceptable use policies applicable" to the GPAI model as per Annexes XI and XII. However, the draft Code goes farther, treating as a requirement that a provider actually *have* an applicable acceptable use policy (AUP) that purports to restrict uses of its GPAI models, and even goes so far as to lay out purported minimum elements required of such a policy.

Not only does this go beyond the intent of the Act, which only requires disclosure of an "applicable" AUP–that is, if there is one–it also poses a distinct hurdle for providers of open source GPAI models. That is because a core feature of most open source licenses–arguably the most important for interoperability between open licenses, a core enabler of a robust open source ecosystem–is that they do

not purport to restrict the uses to which the licensed software or model can be applied.[6] The Act itself clearly contemplates the existence of such licenses by creating an exception for open source GPAI models, eliminating their transparency obligations unless those models also pose systemic risks (see Articles 2(12), 53(2), 54(6z)), based on the recognition that "models…released under a free and open-source licence that allows them to be openly shared and where users can *freely* access, *use*, modify and redistribute them or modified versions thereof, can contribute to research and innovation in the market and can provide significant growth opportunities for the Union economy…[and help] ensure high levels of transparency and openness…." (AI Act recitation 102, emphasis added).

Of course, there has been some evolution in the field of open AI licenses in the past few years such that some licensors have experimented with what are often called Responsible AI Licenses or RAIL licenses that purport to restrict certain harmful or illegal uses. But we do not think the intent of the Act or the drafters was to *require* that all open licenses take this approach, which would be the de facto result of requiring that the provider have AUPs that restrict uses. The drafters should clarify that AUPs are not a requirement for open source model providers, and the Appendix on "Essential elements of an Acceptable Use Policy" should be reframed as describing desirable rather than required attributes of an AUP.

b. The draft Code should provide guidance on the scope of the open source exceptions in the AI Act

As noted above, there are specific exceptions for open source licensing in the AI Act. However, considering the manner in which those exceptions were drafted, and further considering the above-mentioned evolution and experimentation in the realm of open AI licenses, there is real uncertainty about what would count as a "free and open source license" under the Act's terms. As commenters have pointed out, the particular language used in the Act doesn't squarely line up with the features of traditional open source licenses.[7] Yet how these exceptions are interpreted will have an outsized impact on the future of open licensing and the incentives of open AI providers.

For example, it is unclear whether the exception would apply if the license in question included use restrictions such as those found in newer RAIL licenses as discussed above. Yet how this question is answered will likely dictate whether RAIL licenses become a common alternative to more traditional unrestricted-use open licenses. Whether this is a good or bad outcome depends on a policy debate that has not yet occurred overtly, over the question of whether the safety benefits of RAIL licenses–by virtue of enabling providers to issue requests to open model platforms like Hugging Face and Github to take down models that violate the license–outweigh the benefits of traditional licenses including their easy interoperability thanks to the absence of conflicting use restrictions.

Similarly, there are currently debates over whether an AI model is truly "open" if its training data is not also open. For example, version 1.0 of the Open Source Initiative's open source AI definition would

[6] See, e.g., Open Source Initiative, "The Open Source Definition," last modified February 16, 2024, https://opensource.org/osd (including in the open source definition a requirement of "No Discrimination Against Fields of Endeavor. The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a business, or from being used for genetic research."

[7] See, e.g., Kate Downing, "Choose Your Own Adventure: The EU AI Act and Openish AI," February 6, 2024, https://katedowninglaw.com/2024/02/06/choose-your-own-adventure-the-eu-ai-act-and-openish-ai-2/ (extensively analyzing disjunctions between AI Act's language and traditional open source licensing).

require that an open model be accompanied by "[s]ufficiently detailed information about the data used to train the system so that a skilled person can build a substantially equivalent system."[8] Yet such a broad requirement could disincentivise the use of proprietary data in open models, which could in turn suppress the release of open models that can meaningfully compete with closed models that do rely on proprietary data. Again, the question of whether this is an appropriate trade-off is a policy debate that has not yet been had, but is critically important to the future of open source AI.

Especially considering the lack of other venues in which to discuss and provide clarity on these questions, we urge that the drafters use the opportunity of this Code of Practice process to at least begin to address them, including by soliciting views of stakeholders, to help provide greater clarity on what type of open GPAI license will actually satisfy the Act's exception for open source models. Reaching a consensus would be ideal, but even simply framing the questions and highlighting different potential interpretations would be of assistance to open source developers in need of guidance.

c. The draft Code should clarify tool sharing and cybersecurity requirements especially as they apply to open source model providers

Several other provisions of the Code warrant being strengthened or clarified in order to better fulfill the Act's goal of not unduly burdening open source development, including:

Submeasure 10.8. Sharing tools & best practices. We greatly appreciate the goal of this submeasure to help build a more open and accessible ecosystem of "best-in-class safety evaluations, tooling, and accompanying best practices." To better ensure robust contributions from signatories, we would recommend strengthening the language (e.g. from "strive to make…accessible" to "actively contribute to making…accessible"), and to require that those contributions be open source wherever reasonably possible to further ensure their broad accessibility.

Submeasure 12.2. Security mitigations. This submeasure sensibly clarifies that security measures must only be taken to protect access to "unreleased" model weights and other components, not published open source models. However, all open source models were at some point unreleased before they were public–or they were developed in public, collaboratively. In these cases, it is unclear what responsibilities an open source model provider has, if any, to attempt to preserve secrecy around models in development. Requiring substantial technical security around model weights and components intended to be made public would place a substantial burden on open source providers with varying little benefits, and the drafters should further clarify what if any obligations such developers should have under this submeasure.

This final point brings us back to the first point of our submission, about the importance of appropriate prioritisation of systemic risks especially when considering the often limited resources of SMEs and open source projects. Throughout this process, we urge the drafters to ensure that concrete and evidenced risks are prioritised over theoretical or speculative risks so that the limited capacities of GPAI model developers are allocated in a manner most effective at protecting the health, safety, and fundamental rights of Europeans.

[8] Open Source Initiative, "The Open Source AI Definition - v. 1.0," last accessed November 26, 2024, https://opensource.org/ai/open-source-ai-definition.