Deutscher Bundestag
Ausschuss für Digitales

Ausschussdrucksache
20(23)151

23.05.2023

Expert Testimony on Generative AI

Prof. Dr. Philipp Hacker, LL.M. (Yale)*

May 22, 2023

Prepared for:

Öffentliche Anhörung „Generative Künstliche Intelligenz“ am Mittwoch, 24. Mai 2023, 14:30 – 16:30 Uhr, Sitzungssaal Reichstagsgebäude (RTG) 3 N 001

Contents:

* Chair for Law and Ethics of the Digital Society, European New School of Digital Studies, European University Viadrina.

I. General considerations

AI regulation is at a crossroads, both in the EU and beyond.

We are currently witnessing, in real time, the birth of a new generation of AI systems, particularly in the realm of generative AI. These models offer tremendous opportunities and will significantly change–in fact, are already changing–the ways we work, communicate, and live. Simultaneously, this new generation of AI systems harbors specific risks that regulation needs to address. In my view, the most urgent ones, in the short and medium-term, are the following six issues: data protection; non-discrimination; quality (of data and output); content moderation; environmental sustainability; and civil liability. In the longer term, we also need to prepare for a potential restructuring of the job market, with concomitant effects on tax revenue, as well as the use of AI by malicious actors.

As I have pointed out in previous publications,[1] the legal framework for AI, and generative AI in particular, must strike a delicate balance between adequately addressing these risks while allowing AI developers to build models and put them in practice in socially useful ways. Significantly, regulation should foster an ecosystem that avoids further market concentration in the hands of a small number of companies based outside the EU; rather, companies large and small, both outside and inside the EU, must be able to rapidly develop and deploy AI systems within the guardrails of the law. Hence, it is of utmost importance to design regulation in a way that is both effective and operationalizable, i.e., that can–to the best extent possible–be smoothly implemented and complied with by all actors in the AI value chain.

II. Answers to the specific questions prepared by the Committee for Digital Affairs

1. The regulation of generative AI is currently the subject of negotiations on the European AI Act (AIA). In your opinion, how can generative AI be effectively included and regulated in the AIA and how do you assess proposed differentiations within generative AI between "general purpose AI" and "foundation models"?

On May 11, the European Parliament cleared the way for its position on the AI Act with two crucial committee votes. Arguably, the draft is heading into the right direction, but important shortcomings remain that threaten to derail AI development and deployment in the EU.

a) Terminology

Concerning terminology, The EP version of the AI Act introduces a novel provision outlining a specific framework for what it refers to as “foundation models.” This term encompasses

[1] See, e.g., Philipp Hacker, Andreas Engel and Marco Mauer, 'Regulating ChatGPT and other Large Generative AI Models' (2023) ACM Conference on Fairness, Accountability, and Transparency (FAccT '23, forthcoming) https://arxiv.org/abs/2302.02337; Philipp Hacker, 'The European AI Liability Directives - Critique of a Half-Hearted Approach and Lessons for the Future' (2022) Working Paper, https://arxivorg/abs/221113960; see also Philipp Hacker, Andreas Engel and Amelie Berz, The EU AI Act is improving – but still contains fundamental flaws, TechMonitor (May 19, 2023), https://techmonitor.ai/comment-2/eu-ai-act-improving.

highly advanced AI models, including various generative AI systems like ChatGPT, GPT-4, Bard, or Stable Diffusion. The concept of a "foundation model" has gained significant recognition within the computer science community,[2] appropriately emphasizing the broad range of tasks and outputs the models can handle.[3] For instance, a simple classifier solely capable of distinguishing between humans and dogs in images would not meet the criteria, whereas a text generator such as GPT-4 or Luminous, capable of summarizing, completing, and autonomously generating text, would qualify as a foundation model. In my view, it does make sense to use terminology that is also prevalent in the computer science community, and to address foundation models specifically.

By contrast, the term "general-purpose AI system" is inherently vague, used with varying connotations–if at all–in the technical community,[4] and should be abandoned.

b) Regulatory architecture for foundation models: three layers

I refer to our paper, in which we use the term "large generative AI model" (LGAIM) to refer to foundation models.

"[R]egulation of LGAIMs is necessary, but must be better tailored to the concrete risks they entail. Hence, we suggest a shift away from the wholesale AI Act regulation envisioned in the general approach of the Council of EU toward specific regulatory duties and content moderation. Importantly, regulatory compliance must be feasi e for LGAIM developers large and small to avoid a winner-takes-all scenario and further market concentration [82]. This is crucial not only for innovation and or consumer welfare [33, 159, 160], but also for environmental sustainability. While the carbon footprint of IT and AI is significant and steadily rising [54-58], and training of LGAIMs is particularly resource intensive [161], large models may ultimately create fewer greenhouse gas emissions than their smaller brethren if they can be adapted to multiple uses.

Against this background, we envision three layers of obligations for LGAIMs: a first set of minimum standards for all LGAIMs; a second set of specific high-risk rules applying only to LGAIMs used in concrete high-risk use cases; and the third set of rules governing collaboration along the AI value chain (see Section 3.2.2) to enable effective compliance with the first two sets of rules."[5]

i. The May 9 proposal by the European Parliament

(1) First layer: minimum standards for all foundation models

"The first layer will apply to the providers (=developers) of a subset of GPAIS denominated "foundation models" (Article 28b(1)-(3) AI Act EP Version) and generative AI (Article 28b(4) AI Act EP Version). Referring to a well-known term in the computer science community the EP version defines foundation models as an AI system "that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks"

[2] See, e.g., Rishi Bommasani and others, 'On the opportunities and risks of foundation models' (2021) arXiv preprint arXiv:210807258.
[3] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models'.
[4] Carlos Ignacio Gutierrez and others, 'A Proposal for a Definition of General Purpose Artificial Intelligence Systems' (2022) Working Paper, https://ssrncom/abstract=4238951; Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models'.
[5] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 17.

(Article 3(1c) AI Act EP Version) The focus on generality of output and tasks is indeed better suited to capture the specifics of large generative AI models than the vague definition of GPAIS (see Section). In line with suggestions made in this paper, the general obligations for all foundation models include data governance measures, particularly with a view to the mitigation of bias (Article 28b(2)(b) AI Act EP Version). Furthermore, appropriate levels of performance, interpretability, corrigibility, safety and cybersecurity must be maintained throughout the model's lifecycle. These requirements have to be tested for, documented, and verified by independent experts, Article 28b(2)(c) AI Act EP Version. Crucially, however, all foundation models also need to implement risk assessments, risk mitigation measures, and risk management strategies with a view to reasonably foreseeable risks to health, safety, fundamental rights, the environment, democracy and the rule of law, again with the involvement of independent experts, Article 28b(2)(a) AI Act EP Version. Effectively, this requirement is tantamount to classifying foundation models as high-risk per se.

A crucial element of the minimum standards for generative AI is contained in the "ChatGPT Rule" Art. 28b(4) AI Act EP Version. It contains three main elements. (i) The transparency obligation concerning the use of AI (Art. 28b(4) AI Act EP Version, Art. 52(1) AI Act) is a step in the right direction. It addresses obligations of providers towards users of AI systems. In our view, additionally, obligations of users towards recipients are warranted in some instances to fight the spread of fake news and misinformation (see Section). (ii) The rule on preventing a breach of EU law also arguably would benefit from greater detail. Here, the compliance mechanisms of the DSA should be transferred much more specifically, for example through clear, mandatory notice and action procedures and trusted flaggers. It goes without saying that the models must comply with applicable law. (iii) The disclosure of copyrighted material contained in training data may indeed help authors and creators enforce their rights. However, even experts often argue whether certain works are copyrightable at all or not. What must be avoided is that developers who have, e.g., processed 20 million images now have to conduct a full-scale legal due diligence on these 20 million images to decide for themselves whether they are copyrightable or not. Hence, it must therefore be sufficient to disclose, even in an over-inclusive manner, works which may be copyrightable, including those for which it is not clear whether they are ultimately copyrightable or not. Otherwise, again, practically prohibitive due diligence costs will arise. The individual author must then decide, when she discovers her work, whether she thinks it is protected by copyright or not. […]

In our view, while containing steps in the right direction, this proposal would be ultimately unconvincing as it effectively treats foundation models as high-risk applications. Of course, as noted and discussed in detail below, AI output may be misused for harmful speech and acts (as almost any technology). But not only does this seem to be rather the exception than the rule. The argument concerning adverse competitive consequences applies equally here. Under the EP version, risk assessment, mitigation, and management still remain focused on the model itself rather than the use-case specific application (Art. 28b(2)(a) and (f) AI Act EP Version), even though Recital 58a acknowledges that risks related from AI systems can stem from their specific use. Again, this leads to the onerous assessment and mitigation of hypothetical risks that may never materialize–instead of managing risks at the application level where the concrete deployment can be considered."[6]

[6] Ibid, 7 et seq..

(2) Second layer: specific high-risk applications

"The second layer refers to "new providers" which significantly modify the AI system, Art. 28(1)(b) and (ba) AI Act EP Version. This new provider, which is called deployer in our paper, assumes the obligations of the former provider upon substantial modification; the new provider takes on this role (Art. 28(1) and (2)(1) AI Act EP Version).

The new rule on a fundamental rights impact assessment (Article 29a AI Act EP Version) also applies on this level of the concrete application. This rule also seems hardly operationalizable. Fundamental rights are a fuzzy category and difficult to implement at a technical level, where specific secondary regulation may be more useful (GDPR, non-discrimination law directives etc.). Importantly, it is also doctrinally misguided as private companies, in general, cannot violate fundamental rights of other private persons (fundamental rights bind the state, not the citizens; there are only some exceptions to that rule[7])."[8] Furthermore, the relationship of the fundamental rights impact assessment to the general risk assessment (Articles 9 and 28b(1)(a)) is unclear and may lead to unnecessary duplication.

(3) The third layer: collaboration along the AI value chain

"A third layer of requirements relates to the AI value chain (Article 28(2)(2) AI Act EP Version), in line with suggestions made below in this paper."[9]

ii. Own Proposal

In my view, we do need certain rules in the first layer, applying to all foundation models. Not risk assessment and management, but other mechanisms (see below). In the second layer, risk assessment and management should be tied to specific use cases. The third layer needs to provide access and information rights to enable players in the AI value chain to gather the information necessary for compliance with the AI Act.[10]

"Concerning minimum standards, first and foremost, the EU acquis applies to developers of LGAIMs as well, putting the GDPR (see Section 5), non-discrimination law (Section 4),[11] as well as product liability [24] center stage. In addition, transparency rules, now also proposed by the European Parliament [70], must apply (see below, Section 7.1). Furthermore, specific risks of such outstanding relevance that they should be addressed at the upstream level, rather than delegated to deployers in specific use cases, must be allocated to developers as part of the minimum standards. This concerns, in our view, selected data governance duties (Art. 10 AI Act, see Section 4) and rules on the ever more important issue of cybersecurity (Art. 15 AI Act). Finally, sustainability rules [24] as well as content moderation (see below, Section 7.4) should also form part of the minimum standards applicable to all LGAIMs."[12]

---

[7] See, e.g., CJEU, Case C-414/16 (Egenberger) and following judgments in this line.
[8] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 7 et seq..
[9] Ibid, 7 et seq.
[10] On this point more specifically, see ibid, 10 et seq.
[11] See, e.g., Sandra Wachter, 'The Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law' (2022) arXiv preprint arXiv:220501166; Frederik J Zuiderveen Borgesius, 'Strengthening legal protection against discrimination by algorithms and artificial intelligence' (2020) 24 The International Journal of Human Rights 1572; Philipp Hacker, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143.
[12] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 17.

(1) Risk assessment and management: use-case specific

"Importantly, the full extent of the high-risk section of the AI Act, including formal risk management, should only apply if and when a particular LGAIM (or GPAIS) is indeed used for high-risk purposes. This strategy aligns with a general principle of product safety law not every screw and bolt must be manufactured to the highest standards. For example, only if they are used for spaceships, stringent product safety regulations for producing aeronautics material apply[13]–but not if they are sold in the local DIY store for generic use. The same principle should be applied to LGAIMs."[14]

(2) Content moderation

"One of the biggest challenges for LGAIMs is, arguably, their potential misuse for disinformation, manipulation, and harmful speech. In our view, the DSA rules conceived for traditional social networks must be expanded and adapted accordingly. The European Parliament has partially addressed this challenge by stipulating that foundation models must not violate EU law [76]. In our view, however, regulation should go one step further by selectively expanding DSA rules to LGAIM developers and deployers. LGAIMs, and society, would benefit from mandatory notice and action mechanisms, trusted flaggers, and comprehensive audits for models with particularly many users. The regulatory loophole is particularly virulent for LGAIMs offered as standalone software, as is currently the case. In the future, one may expect an increasing integration into platforms of various kinds, such as search engines or social networks, as evidenced by LGAIM development or acquisition by Microsoft, Meta, or Google. While the DSA would then technically apply, it would still have to be updated to ensure that LGAIM-generated content is covered just like user-generated content. In particular, as LGAIM output currently is particularly susceptible to being used for the spread of misinformation, it seems advisable to require LGAIM-generated content to be flagged as such–if technically feasible. Doctrinally, this could be achieved via an amendment of the DSA or of Article 29 AI Act, which already contains notification duties in its para. 4 (see Part. 4). Given the current political process in the EU, the latter option seems more realistic."[15]

(3) Sustainability

„the AI Act should mandate a "sustainability impact assessment" for AI systems. Such an assessment could draw on the manifold proposals concerning impact assessments for AI systems in general.[16] To this end, a provision structurally mirroring Article 9 AI Act (risk management system) should be added to the AI Act. It would apply to developers of high-risk and non-high-risk AI systems alike as the carbon footprint of AI systems is unrelated to their level of risk for health, safety or fundamental rights. Crucially, during the modelling phase, developers should compare different model types (e.g., linear regression versus deep learning) not only with respect to their performance, but also their estimated climate footprint.[17] Already,

[13] See, e.g., product standards, aerospace series, DIN EN 4845–4851 (December 2022) on screws.
[14] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 19.
[15] Ibid, 19 et seq.
[16] Andrew D Selbst, 'An Institutional View of Algorithmic Impact Assessments' (2021) 35 Harvard Journal of Law & Technology; Kaminski and Malgieri, Multi-layered explanations from algorithmic impact assessments in the GDPR; Margot E Kaminski and Gianclaudio Malgieri, 'Algorithmic impact assessments under the GDPR: producing multi-layered explanations' (2020) International Data Privacy Law 19.
[17] The exact impact is not easy to measure. An index including Scope 1, 2 and 3 Emissions for necessary compute resources (e.g., energy; carbon emissions) for training and retraining could be used as a proxy. For a more comprehensive impact measure (including production, transport, and end-of-life, as well as water consumption), see OECD, Measuring the Environmental Impacts of AI Compute and Applications: The AI Footprint, Annex A;

there are tools available to measure the carbon impact of models.[18] Simply put, if two model types exhibit similar performance, the developers would be obliged under the new provision to choose the more sustainable model for further development and deployment. In this way, the current fixation on performance measures can be complemented by even greater environmental awareness and concrete, low maintenance steps to include sustainability in the wider target function of ML development.

In fact, in many scenarios, sustainability and performance may even go hand in hand. A current trend in machine learning is the use of pre-trained models, which have been trained on some more general data for a certain task class (for example, image[19] or speech recognition[20]).[21] They are then fully trained by developers working on a concrete problem with domain-specific data. Such pre-trained models are not only often more performative and have become the state-of-the-art architecture in numerous tasks,[22] but they are also less energy consuming overall as the pre-training only has to be done once for many different model deployments.[23] However, ironically, regulation might thwart these efforts. The most potent pre-trained models are [foundation models]–precisely the ones whose development the current version of AI Act and AI liability directives significantly dis-incentivise. Here, the wheel comes full circle, but not in an efficient or sustainable way. This again points to the importance of changing the rules for"[24] foundation models.

2. Generative AI offers numerous possible applications in a wide range of occupations and can ease the burden on the labor market. How do you assess the potential and risks of generative AI for the world of work and where do you see a need for regulation?

I should note at the outset specialist in labor economics.

However, I firmly believe that generative AI will bring significant change to the labor market in ways both beneficial and detrimental, depending on the affected jobs. These changes cannot be fully prevented, only channeled into certain directions.

Many tasks in the knowledge and creative domain will be supported by generative AI applications that will likely boost productivity. Importantly, generative AI, in many sectors, may also fill the widening gap caused by the overall shortage of skilled laborers.

On the other hand, it is no secret that generative AI will also replace some jobs. At this point, it seems premature to estimate with confidence whether all of these jobs will in the medium

---

on Scope 1, 2 and 3 Emissions, see IPCC, Working Group III Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (2014), 122.

[18] Overview in OECD, Measuring the Environmental Impacts of AI Compute and Applications: The AI Footprint, 28.

[19] See, e.g., Gustavo Carneiro, Jacinto Nascimento and Andrew P Bradley, Unregistered multiview mammogram analysis with pre-trained deep learning models (Springer 2015).

[20] Juliette Millet and others, 'Toward a realistic model of speech processing in the brain with self-supervised learning' (2022) NeurIPS https://arxiv.org/abs/2206.01685.

[21] Xu Han and others, 'Pre-trained models: Past, present and future' (2021) 2 AI Open 225.

[22] Ibid.

[23] David Patterson and others, 'Carbon emissions and large neural network training' (2021) arXiv preprint arXiv:210410350, 15.

[24] Hacker, 'The European AI Liability Directives - Critique of a Half-Hearted Approach and Lessons for the Future', 63 et seq.

term be replaced by other jobs enabled by the rise of generative AI, or if some of them will simply be lost. In my view, prudent regulation should at least make contingency plans for significant losses of jobs in certain sectors particularly affected by the rise of generative AI. Importantly, if a significant number of jobs are lost, this will affect tax revenues. Hence, new models of potentially taxing generative AI applications that do effectively replace jobs, which formally generated taxes, need to be envisioned under such a contingency plan.

It goes without saying that significant investments also need to be made in training and upscaling. Our knowledge-based society will only remain prosperous and productive if we embrace generative AI and become leaders in its development and application.

3. To what extent can applications from state or economic systems that do not always share democratic and liberal values affect European society, and how should the EU and Germany deal with this?

Indeed, arguably the greatest danger emanating from AI is not the model, but humans using it for malicious purposes. One particular problem is the automated mass generation of fake news[25] and hate speech.[26]

"Recent experiments have shown that ChatGPT, despite innate protections [32], may be harnessed to produce hate speech campaigns at scale, including the code needed for maximum proliferation [8]. Furthermore, the speed and syntactical accuracy of LGAIMs make them the perfect tool for the mass creation of highly polished, seemingly fact-loaded, yet deeply twisted fake news [7, 17]. In combination with the factual dismantling of content moderation on platforms such as Twitter, a perfect storm is gathering for the next global election cycle."[27]

The EU, of course, does have a tool designated to stem the tide of illegal content online, particularly with a view to hate speech and certain types of fake news: the Digital Services Act (DSA). However, unfortunately, the DSA does not apply to directly AI systems, including foundation models,[28] creating a dangerous regulatory loophole.

Hence, I suggest extending some of the obligations of the DSA to developers of generative AI. How could this work? "We envision it to have two components. These components would combine centralized and decentralized monitoring within a notice-and-action mechanism (cf. Article 16 DSA).

The first component harnesses the wisdom of the crowd, as it were, to correct LGAIM output. Users should be enabled to flag problematic content and give notice. A special status should be given to a specific group of users, trusted flaggers (cf. Article 22 DSA), who could be private individuals, technologies savvy NGOs, or volunteer coders. After registering with the competent authority, they would essentially function as a decentralized content monitoring team. They could experiment with different prompts and see if they manage to generate harmful or otherwise problematic content. They could also scan the internet for tools to circumvent

[25] Oxford Analytica, 'Generative AI carries serious online risks' (2023) Emerald Expert Briefings .
[26] See, e.g., Chris Stokel-Walker and Richard Van Noorden, 'What ChatGPT and generative AI mean for science' (2023) 614 Nature 214.
[27] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 2.
[28] This is due to the fact that companies developing or deploying foundation/generative AI models are not intermediaries hosting other persons' content, but creating content themselves; see ibid, Part 6.

content moderation policies and instruments at LGAIMs. If they find something, trusted flaggers would send a notice containing the prompt and the output to a content moderation check-in point of the respective LGAIM system, which would forward the notice to developers and/or deployers.

Here, the second component enters the scene, geared toward tech engineers working with developers or deployers. They would have to respond to notices; those submitted by trusted flaggers would have to be prioritized by the content moderation team. Their job, essentially, is to modify the AI system, or to block its output, so that the flagged prompt does not generate problematic output anymore, and to generally search for ways to block easy workarounds likely tried by malicious actors. Furthermore, if the LGAIM system is large enough, they would be tasked with establishing a more comprehensive compliance system (cf. Article 34-35 DSA). Overall, such a combination of centralized and decentralized monitoring could prove more effective and efficient than current systems relying essentially on goodwill to handle the expected flood of hate speech, fake news and other problematic content generated by LGAIMs."[29]

4. So far, there have been a number of ideas and projects ranging from watermarks to tools that are supposed to mark or recognize AI-generated texts - both of which are being critically commented on in view of their lack of consistency or accuracy. What could a secure and effective way of making content created by generative AI known look like in concrete terms? And what accompanying information could be provided to users for the purpose of education?

Watermarks are indeed a good idea, but might in fact prove too easy to eliminate in a number of cases. In terms of transparency, I would make the following suggestions:

a) EP Proposal

"(i) The transparency obligation concerning the use of AI (Art. 28b(4) AI Act EP Version, Art. 52(1) AI Act) is a step in the right direction. It addresses obligations of providers towards users of AI systems. In our view, additionally, obligations of users towards recipients are warranted in some instances to fight the spread of fake news and misinformation."[30]

b) Obligations for developers and deployers

"First, LGAIM developers and deployers should be required to report on the provenance and curation of the training data, the model's performance metrics, and any incidents and mitigation strategies concerning harmful content. Ideally, to the extent technically feasible [54, p. 28, Annex A], they should also disclose the model's greenhouse gas (GHG) emissions, to allow for comparison and analysis by regulatory agencies, watchdog organizations, and other interested parties. This information could also serve as the basis for an AI Sustainability Impact Assessment."[31]

[29] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 20.
[30] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 7.
[31] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 18.

c) Transparency obligations for users

"Second, *professional users* should be obligated to disclose which parts of their publicly available content were generated by LGAIMs, or adapted based on their output. Specifically, this entails that in adidas example, adidas needs to adequately inform users that the design was generated using, e.g., Stable Diffusion. While the added value of such information may be limited in sales cases, such information is arguably crucial in any cases involving content in the realm of journalism, academic research, or education. Here, the recipients will benefit from insight into generation pipeline. They may use such a disclosure as a warning signal and engage in additional fact checking or to at least take the content *cum grano salis*. Eventually, we imagine differentiating between specific use cases in which AI output transparency vis-à-vis recipients is warranted (e.g., journalism, academic research or education) and others where, based on further analysis and market scrutiny, such disclosures may not be warranted (certain sales, production and B2B scenarios, for example). For the time being, however, we would advocate a general disclosure obligation for professional users to generate further information and insight into the reception of such disclosures by other market participants or recipients.

Conversely, we submit that *non-professional users* should not be required to inform about the use of AI. In the birthday example, hence, a parent would not need to inform the parents that the invitation or the entire design of the birthday party was rendered possible by, e.g., Aleph Alpha's Luminous or ChatGPT. One might push back against this in cases involving the private use of social media, particularly harmful content generated with the help of LGAIMs. However, any rule to disclose AI-generated content would likely be disregarded by malicious actors seeking to post harmful content. Eventually, however, one might consider including social media scenarios into the domain of application of the transparency rule if AI detection tools are sufficiently reliable. In these cases, malicious posts could be uncovered, and actors would face not only the traditional civil and criminal charges, but additionally AI Act enforcement, which could be financially significant (administrative fines) and hence create even greater incentives to comply with the transparency rule, or refrain from harmful content propagation.

The enforcement of any user-focused transparency rule being arduous, it must be supported by technical measures such as digital rights management and watermarks imprinted by the model The European Parliament is currently pondering a watermark obligation for generative AI Importantly, more interdisciplinary research is necessary to develop markings that are easy to use and recognize, but hard to remove by average users This should be coupled with research on AI-content detection to highlight such output where watermarks fail".[32]

5. Currently, numerous proposals are circulating to accurately anchor the regulatory challenges of generative AI applications in the EU legislative projects for an AI Regulation and an AI Liability Directive: Is the risk-based approach to regulating generative AI suitable at all or do we need, for example, a systemic risk analysis analogous to the risk analysis and minimization mechanism in the DSA?

My answer to this would be twofold. On the one hand, as a society and as researchers, we should embark on systemic risk analysis of foundation models, contemplating also the possible long-term effects in terms of employment, tax revenue, and use by malicious actors.

[32] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 18.

However, I firmly believe that, with respect to the AI Act and the AI liability directives, regulation should adhere to a use-case-specific architecture concerning risk assessment, mitigation and management. Comprehensive and systemic risk analysis might be undertaken by the AI office, national regulators, or dedicated NGOs.

AI developers are arguably not well-placed to conduct such an assessment. "Setting up a [comprehensive systemic risk management] system seems to border on the impossible, given LGAIMs' versatility. [Under the Councils general approach,] it would compel LGAIM providers to identify and analyze all "known and foreseeable risks most likely to occur to health, safety and fundamental rights" concerning all possible high-risk uses of the LGAIM (Articles 9(2)(a), 4b(6) AI Act Council Version). On this basis, mitigation strategies for all these risks have to be developed and implemented (Article 9(2)(d) and (4) AI Act Council Version). Providers of LGAIMs such as ChatGPT would, therefore have to analyze the risks for every single, possible application in every single high-risk case contained in Annexes II and III concerning health, safety and all possible fundamental rights.

Similarly, performance, robustness, and cybersecurity tests will have to be conducted concerning all possible high-risk uses (Articles 15(1), 4b(6) AI Act Council Version). This seems not only almost prohibitively costly but also hardly feasible. The entire analysis would have to be based on an abstract, hypothetical investigation, and coupled with–again hypothetical–risk mitigation measures that will, in many cases, depend on the concrete deployment, which by definition has not been implemented at the moment of analysis. What is more, many of these possible use cases will, in the end, not even be realized because they are economically, politically, or strategically unviable. Hence, such a rule would likely create "much ado about nothing", in other words: a waste of resources. Ironically, the conception of Articles 4a-4c AI Act, as currently proposed, places a very high, and arguably undue, burden on providers of truly general-purpose AI systems. These providers will be most unlikely to be able to comply with the AI Act, by virtue of their model's sheer versatility–there will just be too many scenarios to contemplate. In conjunction with the proposed regime for AI liability, which facilitates claims for damages if the AI Act is breached, this also exposes LGAIM providers to significant liability risk […].

[Such] rules would likely have significantly adverse consequences for the competitive environment surrounding LGAIMs. The AI Act definition specifically includes open source developers as LGAIM providers, of which there are several.[33] Some of these will explore LGAIMs not for commercial, but for philanthropic or research reasons. For example, Stable Diffusion was developed in a research project conducted at LMU Munich. While, according to its Article 2(7), the AI Act shall not apply to any (scientific, see Recital 12b AI Act) research and development activity regarding AI systems, this research exemption arguably does not apply anymore once the system is released into the wild, as any public release likely does not have scientific research and development as its "sole purpose" (Recital 12b AI Act), particularly when, as is often the case, a commercial partner enters to limit liability and provide necessary fine-tuning.

As a result, all entities–large or small–developing LGAIMs and placing them on the market will have to comply with the same stringent high-risk obligations. Given the difficulty to comply with them, it can be expected that only large, deep-pocketed players (such as Google,

[33] See, e.g., https://www.kdnuggets.com/2022/09/john-snow-top-open-source-large-language-models.html.

Meta, Microsoft/Open AI) may field the costs to release an approximately AI Act-compliant LGAIM. For open source developers and many SMEs, compliance will likely be prohibitively costly. Hence, the AI Act may have the unintended consequence of spurring further anti-competitive concentration in the LGAIM development market. This is in direct opposition to the spirit of Recital 61 Sentence 5 AI Act which–in the context of standardization–explicitly calls for an appropriate involvement of SMEs to promote innovation and competitiveness in the field of AI within the Union (see also Article 40(2)(b) and Article 53(1b)(a) AI Act). Similar effects have already been established concerning the GDPR In this sense, the AI Act threatens to undermine the efforts of the Digital Markets Act[34] to infuse workable competition into the core of the digital and platform economy."[35]

6. **Are new phenomena and issues to be expected with regard to a negative influence of applications of generative AI on the democratic opinion-forming process, and how can media freedom and diversity of opinion be legally and politically strengthened in the age of generative AI, also - but not exclusively - with regard to the appropriate remuneration of journalists, artists and creatives, and where do you see a possible need for adaptation, for example in copyright law?**

Concerning the democratic processes and public opinion, see my answer to Question 3.

Concerning copyright law, the EU already has a mechanism in place with Art. 3 and 4 C-DSM Directive. Here, most urgently, the veto right afforded to rightholders under Article 4(3) C-DSM must be standardized so that AI developers are in a position to evaluate whether the works they intend to train the model on can be used for non-research AI training.

As for remuneration, I do not have any specific insights to share on that point.

7. **What legal starting points are there in EU law (e.g. AI Act, competition law, Copyright Directive) and in national law (e.g. UWG, State Media Treaty) to implement a labeling obligation for AI-generated content (e.g. videos, images or texts) and decisions, if possible without circumvention possibilities - and what technical starting points are conceivable to effectively implement and enforce such obligations in digital services?**

At the EU level, the logical place to implement this would be the AI Act. However, this would mean that the rule would likely only be binding from 2024 or 2025 on. Hence, it would be advisable to implement such a rule rapidly in national law. The State Media Treaty (Medienstaatsvertrag) could be a good starting point here, or simply a novel federal legal act that would, at some point, be superseded by the EU AI Act.

[34] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector, OJ L265/1 (DMA).
[35] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 5-6.

8. **What technical and organizational measures do you consider suitable for the protection of minors - both with regard to the inclusion of their personal data in the training and learning environment of generative AI and with regard to the actual use of applications that generate AI-based texts, videos or images?**

This is an excellent question. I would suggest the following architecture:

- age verification tools, as apparently implemented by OpenAI in response to the requirements by the Italian Data Protection Authority.
- mandatory and minor-specific content moderation tools, aligned with the BIK+ strategy,[36] triggered by specific age limits
    - Three core points of BIK+:
        - "Safe digital experiences, protecting children from harmful and illegal online content, conduct, and risks and improving their well-being through a safe, age-appropriate digital environment.
        - Digital empowerment so that children acquire the necessary skills and competencies to make informed choices and express themselves in the online environment safely and responsibly.
        - Active participation, respecting children by giving them a say in the digital environment, with more child-led activities to foster innovative and creative safe digital experiences."
- personal data of children and adolescents is already protected under the GDPR.
- furthermore, Article 28b of the Audiovisual Media Services Directive must be consistently enforced.

9. **What AI-driven economic development do you forecast for the German and European economies in the short, medium and long term in view of their respective specific structure and do you assume a positive or negative development with regard to the implications for the real economic performance of these economies in a global comparison, also depending on regulation?**

I am not an expert on these matters.

10. **What is your opinion of the letter from the Future of Life Institute signed by many recognized AI experts: To what extent do you share the concerns expressed in it and do you consider the demands formulated in it to be reasonable?**

Regulation should focus on the immediate risks of AI (data protection, discrimination, liability, sustainability, content moderation, quality). We should, of course, not close our eyes to long-term risks. However, the proposed moratorium does not help in this sense.

[36] https://digital-strategy.ec.europa.eu/en/policies/strategy-better-internet-kids.

Generally, I think a moratorium can be a good idea, but only if a) it really involves all relevant players and b) it is likely that the duration of the moratorium will allow for meaningful regulation. Given the current geopolitical situation with Russia and China, and fierce competition in the "West", I do not think that a) is realistic; given the disagreement with GenAI regulation, in the EU and beyond, b) does not seem an option, either. Particularly not in 6 months.

Even if b) was taken for granted–for the sake of analysis–we would have to think hard about what a "Western-led" moratorium would entail in terms of AI development (by noncomplying states and actors); deployment of these unsafe systems–both in developing countries and in the Western world–; geopolitical strategy; and competitiveness of research and industry in the EU and the US. These are all hard questions.

Politically speaking, I do not think a mandatory moratorium is a realistic option in the EU or the US.

11. According to the German AI Association, investments of 300 million euros are needed to expand a computing infrastructure in Germany for training algorithms. In your opinion, should it be the task of the state to pursue an active industrial policy by (co-)financing such an infrastructure in order to enable German companies to survive in the global market for generative AI?

Yes, and €300 million are very likely not even enough to match the investments made in the US and the UK, for example. The state must assume a leading role in the endeavor to facilitate a European compute infrastructure, preferably together with other EU states. The market is already concentrated and many major developments take place outside of the EU. Arguably, market shares are allocated right now. If you are is supposed to compete in this race–which it must due to geopolitical sovereignty and independence–, we desperately need these resources.

12. There is widespread agreement to regulate artificial intelligence in such a way that its use follows certain value concepts. How can this be realized in concrete terms and where should the line be drawn with regard to possible overregulation, in which artificial intelligence could become artificial ideology?

To draw this line will always be tricky in individual cases, as the jurisprudence on harmful speech shows. However, the answer, I believe, must be regulation for trustworthy AI and content moderation of generative AI, as sketched above (Questions 3 and 6).

13. So far, almost three quarters of all large AI foundation models come from the USA, and another fifteen percent from China. Against this background, what measures should policymakers in Germany and Europe take as a matter of priority with a view to promoting and strengthening the ecosystem of generative AI if we want to avoid becoming completely dependent on non-European foundation models and only being able to act as purchasers of these models at the end of the value chain?

See the answer to Question 11.

Furthermore, we must pave the way to attract and retain international talent in the AI space.

Finally, we may need specific support of European SMEs, both in financial and regulatory terms. For example, we could help European SMEs with preferential access to sandboxes; more lenient provisions concerning the implementation of the AI Act requirements; and financial support concerning insurance.

14. In your view, what rules are needed in the AI Act for generative AI, specifically with regard to the obligations for developers of foundation models to pass on information within the supply chain, what advantages and disadvantages are associated with such obligations, and at what threshold should the high-risk rules provided for in the AI Act apply to applications based on generative AI?

See the answer to Question 1.

Concerning information and access rights in the AI value chain more specifically:

"individual actors in the AI value chain may simply not have the all-encompassing knowledge and control that would be required if they were the sole addressees of regulatory duties This more abstract observation also shows that shared and overlapping responsibilities may be needed.

In our view, the only way forward are collaborations between LGAIM providers, deployers and users with respect to the fulfillment of regulatory duties, where the regulator gives this (forced) collaboration adequate contours. More specifically, we suggest a combination of strategies known from pre-trial discovery, trade secrets law, and the GDPR. Under the current Council GA AI Act, such teamwork is encouraged in Article 4b(5): providers "shall" cooperate with and provide necessary information to users. A key issue, also mentioned in the Article, is access to information potentially protected as trade secrets or intellectual property (IP) rights In this regard, Article 70(1) AI Act requires anyone "involved" in the application of the AI Act to "put appropriate technical and organizational measures in place to ensure the confidentiality of information and data obtained in carrying out their tasks and activities". To be workable, this obligation needs further concretization; the same holds true for the proposal by the European Parliament in this direction Art. 10(6a) AI Act EP Version only explicitly addresses a situation where such cooperation does not take place, and is limited to violations of Art. 10.

The problem of balancing collaboration and disclosure with the protection of information is not limited to the AI Act. In our view, it has an internal and external dimension. Internally, i.e., in the relationship between the party requesting and the party granting access, access rights are

often countered, by the granting party, by reference to supposedly unsurmountable trade secrets or IP rights The liability directives proposed by the EU Commission, for example, contain elaborate evidence disclosure rules pitting the compensation interests of injured persons against the secrecy interests of AI developers and deployers Article 15(4) GDPR contains a similar provision, which by way of analogy also applies to the access right in Article 15(1) GDPR

Extensive literature and practical experience concerning this problem exists in the realm of the US pretrial discovery system Under this mechanism, partially adopted by the proposed EU evidence disclosure rules injured persons may seek access to documents and information held by the potential defendant before even launching litigation. This, in turn, may lead to non-meritorious access requests by competitors. Such concerns are not negligible in the AI value chain. Here as well, developers, deployers and users may indeed not only be business partners but also be (potential) competitors. Hence, deployers' and users' access must be limited. Conversely, some flow of information must be rendered possible to operationalize compliance with high-risk obligations by deployers.

To guard against abuse, we suggest a range of measures. On the one hand, providers (and potentially deployers) may authorize the use of the model under the proviso that users sign *NDAs* and non-compete clauses. Private ordering should, to a certain extent, function between professional actors. On the other hand, it may be worthwhile to introduce provisions inspired by the US pretrial discovery system and the proposed EU evidence disclosure mechanism (Article 3(4) AI Liability Directive, protective order). Hence, courts should be empowered to issue *protective orders*, which endow nondisclosure agreements with further weight and subject them to potential administrative penalties. The order may also exempt certain trade secrets from disclosure or allow access only under certain conditions (see F.R.C.P. Rule 26(c)(1)(G)). Furthermore, as the high-profile document review cases in the US concerning former and current US Presidents show, the appointment of a *special master* may, ultimately, strike a balance between information access and the undue appropriation of competitive advantage (cf. F.R.C.P. Rule 53(a)) With these safeguards in place, LGAIM developers should be compelled, and not merely encouraged, to cooperate with deployers and users concerning AI Act compliance if they have authorized the deployment.

Concerning the external dimension, the question arises of who should be responsible for fulfilling pertinent duties and be ultimately liable, regarding administrative fines and civil damages, if high-risk rules are violated. Here, we may draw inspiration from Article 26 GDPR (see also. According to this provision, joint data controllers may internally agree on the bespoke allocation of GDPR duties (Article 26(1) GDPR), but remain jointly and severally liable (Article 26(3) GDPR). The reason for this rule is to facilitate data subjects' compensation, who must not fear to be turned away by both controllers with each blaming the other party. Moreover, the essence of the internal compliance allocation must be disclosed (Article 26(2) GDPR). This mechanism could, mutatis mutandis, be transferred to the AI value chain. Here again, collaboration is required and should be documented in writing to facilitate ex post accountability. Disclosing the core parts of the document, sparing trade secrets, should help potential plaintiffs choosing the right party for any ensuing disclosure of evidence requests under the AI liability regime. Finally, joint and several liability ensures collaboration and serves the compensation interests of injured persons. Internally, parties held liable by injured persons can then turn around and seek reimbursement from others in the AI value chain. For example, if the developers essentially retain control via an API distribution model, the internal liability burden will often fall on them. Developers' and deployers' liability, however, must end where

their influence over the deployed model ends. Beyond this point, only the users should be the subject of regulation and civil liability (and vice versa, for example in control-via-API cases): incentives for action only make sense where the person incentivized is actually in a position to act In the GDPR setting, this was effectively decided by the CJEU in the Fashion ID case (CJEU, C-40/17, para. 85). The sole responsibility of the users for certain areas should then also be included in the disclosed agreement to inform potential plaintiffs and foreclose non-meritorious claims against the developer and deployer. Such a system, in our view, would strike an adequate balance of interests and power between LGAIM developers, deployers, users, and affected persons.

Overall, the EP version of the AI Act now rightly contains rules on the AI value chain However, these need to be rendered more specific, as laid out in the preceding sections, to function effectively. Ultimately, allocating responsibility and liability along the value chain is crucial if the AI Act seeks to maintain its spirit of a technology-specific instrument that does not, however, regulate models per se, but primarily models in concrete use cases."[37]

15. What initiatives are there, especially for large language models (LLMs), for the development of European models and how do you assess the possibilities and limitations of Private Public Partnerships in this area?

Concerning concrete models and companies, the following seem important to me:

- Aleph Alpha
- Mistral
- Bloom

Public-private partnerships can be helpful, I suppose, if the state co-finances certain crucial parts of the infrastructure, see the answer to Question 11.

16. In your estimation, what are the next development stages of generative AI, after language and video models (key points AI agents, embodied AI, etc.) and where do the greatest opportunities for our society and economy lie here?

Others are better placed to answer this question.

[37] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 10-11.

17. To what extent does the distribution of advantages and disadvantages through GPAI differ between different population groups (both within national societies and viewed global with a view to the global South/North) due to the aspects listed below:- Differences in access to technology (e.g., due to differences in technical, material, educational, and other prerequisites); Different representation in training data (e.g., health data of women vs. men, of whites vs. PoC, African languages vs. English, etc.); Differential exposure to stereotypical attributions and discrimination (e.g., based on gender or ethnicity); Differential burden of resource consumption caused by AI systems; and how would a more equitable distribution of advantages and disadvantages be achievable?

"Furthermore, we suggest that, as an exception to the focus on LGAIM deployers, certain data curation duties, for example representativeness and approximate balance between protected groups (cf. Article 10 AI Act), should apply to LGAIM developers. Discrimination, arguably, is too important a risk to be delegated to the user stage and must be tackled during development and deployment. Wherever possible, discrimination AI systems should be addressed at its roots (often the training data) and not propagated down the ML pipeline or AI value chain. After all, discriminatory output should, in our view, be avoided in all use cases, even on birthday cards. The regulatory burden, however, must be adapted to the abstract risk level and the compliance capacities (i.e., typically the size) of the company. For example, LGAIM developers should have to pro-actively audit the training data set for misrepresentations of protected groups, in ways proportionate to their size and the type of training material (curated data vs. Twitter feeds scraped from the Internet), and implement feasible mitigation measures. At the very least, real-world training data ought to be complemented with synthetic data to balance historical and societal biases contained in online sources. For example, content concerning professions historically reserved for one gender (nurse; doctor) could be automatically copied and any female first names or images exchanged by male ones, and vice versa, creating a training corpus with more gender-neutral professions for text and image generation."[38]

18. Should generative AI as multi-purpose AI in principle be classified as high-risk AI in the sense of the European AI Regulation in order to meet higher standards and how sensible/feasible do you consider regulatory options for generative AI such as transparency obligations on training data and training processes, the obligation for risk assessment by providers of a GPAI and its publication, visible or invisible labeling of all or certain AI-generated content, the right to verifiability of non-discrimination and access for researchers: Inside and other options discussed?

I think it would be plain wrong to classify GPAIS/foundation models/generative AI as high-risk.[39] See Question 1.

Rather, the architecture of the AI Act is to establish high-risk obligations for specific high-risk use cases. This architecture should not be turned on its head for generative AI.

[38] Hacker, Engel and Mauer, 'Regulating ChatGPT and other Large Generative AI Models', 19.
[39] See also Natali Helberger and Nicholas Diakopoulos, 'ChatGPT and the AI Act' (2023) 12 Internet Policy Review.

III. Most important References

- Bommasani R and others, 'On the opportunities and risks of foundation models' (2021) arXiv preprint arXiv:210807258
- Edwards L, 'Regulating AI in Europe: four problems and four solutions' (2022) 2022
- Geradin D, Karanikioti T and Katsifis D, 'GDPR Myopia: how a well-intended regulation ended up favouring large online platforms' (2021) 17 European Competition Journal 47
- Grinbaum A and Adomaitis L, 'The Ethical Need for Watermarks in Machine-Generated Language' (2022) arXiv preprint arXiv:220903118
- Gutierrez CI and others, 'A Proposal for a Definition of General Purpose Artificial Intelligence Systems' (2022) Working Paper, https://ssrncom/abstract=4238951
- Hacker P, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143
- Hacker P, 'The European AI Liability Directives - Critique of a Half-Hearted Approach and Lessons for the Future' (2022) Working Paper, https://arxivorg/abs/221113960
- Hacker P, Engel A and Mauer M, 'Regulating ChatGPT and other Large Generative AI Models' (2023) ACM Conference on Fairness, Accountability, and Transparency (FAccT '23, forthcoming) https://arxiv.org/abs/2302.02337
- Hacker P and Passoth J-H, 'Varieties of AI Explanations under the Law. From the GDPR to the AIA, and Beyond' (2022) International Conference on Extending Explainable AI Beyond Deep Models and Classifiers 343
- Helberger N and Diakopoulos N, 'ChatGPT and the AI Act' (2023) 12 Internet Policy Review
- Kiela D and others, 'The hateful memes challenge: Detecting hate speech in multimodal memes' (2020) 33 Advances in Neural Information Processing Systems 2611
- Kirchenbauer J and others, 'A Watermark for Large Language Models' (2023) arXiv preprint arXiv:230110226
- Liang P and others, 'The time is now to develop community norms for the release of foundation models' CRFM <https://crfm.stanford.edu/2022/05/17/community-norms.html> accessed February 3, 2023
- Mitchell E and others, 'DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature' (2023) arXiv preprint arXiv:230111305
- Stokel-Walker C and Van Noorden R, 'What ChatGPT and generative AI mean for science' (2023) 614 Nature 214
- Wachter S, 'The Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law' (2022) arXiv preprint arXiv:220501166
- Zuiderveen Borgesius FJ, 'Strengthening legal protection against discrimination by algorithms and artificial intelligence' (2020) 24 The International Journal of Human Rights 1572