Regulating ChatGPT and other Large Generative AI Models

Working Paper, this version February 7, 2023

Philipp Hacker

European New School of Digital Studies, European University Viadrina, hacker@europa-uni.de

Andreas Engel

Heidelberg University, andreas.engel@igw.uni-heidelberg.de

Marco Mauer

Humboldt University of Berlin, marco.mauer@hu-berlin.de

Abstract:

Large generative AI models (LGAIMs), such as ChatGPT or Stable Diffusion, are rapidly transforming the way we communicate, illustrate, and create. However, AI regulation, in the EU and beyond, has primarily focused on conventional AI models, not LGAIMs. This paper will situate these new generative models in the current debate on trustworthy AI regulation, and ask how the law can be tailored to their capabilities. After laying technical foundations, the legal part of the paper proceeds in four steps, covering (1) direct regulation, (2) data protection, (3) content moderation, and (4) policy proposals. It suggests a novel terminology to capture the AI value chain in LGAIM settings by differentiating between LGAIM developers, deployers, professional and non-professional users, as well as recipients of LGAIM output. We tailor regulatory duties to these different actors along the value chain and suggest four strategies to ensure that LGAIMs are trustworthy and deployed for the benefit of society at large. Rules in the AI Act and other direct regulation must match the specificities of pre-trained models. In particular, regulation should focus on concrete high-risk applications, and not the pre-trained model itself, and should include (i) obligations regarding transparency and (ii) risk management. Non-discrimination provisions (iii) may, however, apply to LGAIM developers. Lastly, (iv) the core of the DSA's content moderation rules should be expanded to cover LGAIMs. This includes notice and action mechanisms, and trusted flaggers. In all areas, regulators and lawmakers need to act fast to keep track with the dynamics of ChatGPT et al.

CCS CONCEPTS •Social and professional topics~Computing / technology policy~Government / technology policy~Governmental regulations •

Additional Keywords and Phrases: LGAIMs, LGAIM regulation, general-purpose AI systems, GPAIS, foundation models, large language models, LLMs, AI regulation, AI Act, direct AI regulation, data protection, GDPR, Digital Services Act, content moderation, non-discrimination laws, policy proposals, high-risk applications, transparency obligations, notice and action mechanisms, trusted flaggers

ACM Reference Format

1 INTRODUCTION

Large generative AI models (LGAIMs) are rapidly transforming the way we communicate, illustrate, and create. Their consequences are bound to affect all sectors of society, from business development to medicine, from education to research, and from coding to the arts. LGAIMs harbor great potential, but also carry significant risk. Today, they are relied upon by millions of users to generate human-level text (e.g., ChatGPT), images (e.g., Stable Diffusion, DALL□E 2), videos (e.g., Synthesia), or audio (e.g., MusicLM). Soon, they may be part of employment tools ranking and replying to job candidates, or of hospital administration systems drafting letters to patients based on case files. Freeing up time for professionals to focus on substantive matters–for example, actual patient treatment–, such multi-modal decision engines may contribute to a more effective, and more just, allocation of resources. However, errors will be costly, and risks need to be adequately addressed [1-3]. Already now, LGAIMs' unbridled capacities may be harnessed to take manipulation, fake news, and harmful speech to an entirely new level [4-7]. As a result, the debate on how (not) to regulate LGAIMs is becoming increasingly intense [8-17].

In this paper, we argue that regulation, and EU regulation in particular, is not only ill-prepared for the advent of this new generation of AI models, but also sets the wrong focus by quarreling mainly about direct regulation in the AI Act at the expense of the, arguably, more pressing content moderation concerns under the Digital Services Act (DSA).

Significantly, the EU is spearheading efforts to effectively regulate AI systems, with specific instruments (AI Act, AI Liability Directive), software regulation (Product Liability Directive) and acts addressed toward platforms, yet covering AI (Digital Services Act; Digital Markets Act). Besides, technology-neutral laws, such as non-discrimination law, and also data protection law, continue to apply to AI systems. As we shall see, it may be precisely their technology-agnostic features that make them better prepared to handle the specific risks of LGAIMs than the specific AI regulation that has been enacted or is in preparation.

AI regulation, in the EU and beyond, has primarily focused on conventional AI models, however, not on the new generation whose birth we are witnessing today. The paper will situate these new generative models in the current debate on trustworthy AI regulation, and ask what novel tools might be needed to tailor current and future law to their capabilities. Also, we suggest that the terminology and obligations in the AI Act and other pertaining regulation be further differentiated to better capture the realities of the evolving AI value chain.

To do so, the paper proceeds in five steps. First, we cover technical foundations of LGAIMs, and typical scenarios of their use, to the extent that they are necessary for the ensuing legal discussion. Second, we critique the EU AI Act, which seeks to directly address risks by AI systems. The version adopted by the Council contains provisions to explicitly regulate LGAIMs, even if their providers are based outside of the EU (Art. 4a-c AI Act[1]) [10, cf. also 18]. These proposals, currently hotly debated in the European Parliament [19], arguably fail to accommodate the capacities and broad applicability of LGAIMs, particularly concerning the obligation for an encompassing risk management system covering all possible high
risk purposes (Art. 9 AI Act) [8, pp. 6-10, 20, pp. 13, 51 et seqq.]. Precisely because LGAIMs are so versatile, detailing and mitigating every imaginable high-risk use seems both prohibitive and unnecessary. LGAIM risk regulation should generally focus on deployed applications, not the pre-trained model [8, 20]. However, non-discrimination provisions may apply more broadly to the pre-trained model itself to mitigate bias at its data source. Third, we highlight key data protection
risks under the GDPR, with a particular focus on model inversion [21-23].

Fourth, we turn to content moderation [see, e.g., 24, 25, 26]. Recent experiments have shown that ChatGPT, despite innate protections [27], may be harnessed to produce hate speech campaigns at scale, including the code needed for

[1] Unless otherwise noted, all references to the AI Act are to the general approach adopted by the EU Council on Dec. 6, 2022.

maximum proliferation [5]. 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 [4, 13]. 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. We show that the EU's prime instrument to combat harmful speech, the Digital Services Act (DSA) [28, 29], does not apply to LGAIMs, creating a dangerous regulatory loophole. The paper finishes by making four distinct policy proposals to ensure that LGAIMs are trustworthy and deployed for the benefit of society at large: direct regulation of LGAIM deployers and users, including (i) transparency and (ii) risk management; (iii) the application of non discrimination provisions to LGAIM developers; and (iv) specific content moderation rules for LGAIMs.

2 TECHNICAL FOUNDATIONS OF LARGE GENERATIVE AI MODELS AND EXEMPLARY USAGE SCENARIOS

The AI models covered by this article are often referred to as 'foundation models' [30], 'large language models' (LLMs) [31] or 'large generative models' (LGAIMs – the term adopted in this article) [32]. Although the emergence of

these models in recent years constitutes a significant technical advance, they harness, to great extent, existing technologies in a vastly increased scale and scope. LGAIMs are usually trained with several billion, if not hundreds of billions, parameters [33, 34]. Their training requires large amounts of training data and computing power [35]. OpenAI's "CLIP" image classifier, for example, was built using a set of 400 million image-text pairs [36]. The 'BASIC' model even uses 6.6 billion such pairs [37]. The energy required to train models this large has triggered concerns from a climate policy perspective [20, 38-42].

Hence, LGAIMs "*are advanced machine learning models that are trained to generate new data, such as text, images, or audio*" (Prompt 1, see Annex H1). This "*makes them distinct from other AI models* [*… only*] *designed to make predictions or classifications*" (Prompt 2) or to fulfil other specific functions. This increased scope of application is one of the reasons for the large amount of data and compute required to train them. LGAIMs employ a variety of techniques [27] that aim at allowing them "*to find patterns and relationships in the data on its own, without being* [*explicitly*] *told what to look for. Once the model has learned these patterns, it can generate new examples that are similar to the training data*" (Prompt 3). In simple terms, training data are represented as probability distributions. By sampling and mixing from them, the model can generate content beyond the training data set–thus something new [43, 44]. LGAIMs can often use human text input [45] and produce an output (text; image; audio; video) based on it. The vast amounts of data required imply that developers of LGAIMs often must rely on training data that is openly available on the internet, which can hardly be considered perfect from a data quality perspective [46]. The content generated by these models can, therefore, be biased and prejudiced [11, 47]. To avoid or at least mitigate this issue, model developers need to use proper curating techniques [48].

ChatGPT itself sums up the problem of having to curate its training data and moderate its output this way: "[*T]he models are designed to generate new content that is similar to the training data, which may include offensive or inappropriate content.* [...] *Furthermore, large generative models can generate synthetic content that is difficult to distinguish from real content, making it challenging to differentiate between real and fake information. […* T]*he sheer volume of content generated by these models can make it difficult to manually review and moderate all of the generated content*" (Prompt 4). For as much as we know [27], and according to ChatGPT itself, the creators of ChatGPT sought to address this problem by using "*a combination of techniques to detect and remove inappropriate content. This process includes pre-moderation, where a team of human moderators review and approve content before it is made publicly available. Additionally, ChatGPT uses filtering, which involves using natural language processing and machine learning*

algorithms to detect and remove offensive or inappropriate content. This is done by training a machine learning model on a dataset of examples of inappropriate content, and then using this model to identify similar content in new inputs" (Prompt 5). While we cannot perfectly verify these claims due to lack of transparency on OpenAI's side, it seems that ChatGPT relied or relies on humans that train an automatic content moderation system to prevent the output from becoming abusive [49].

Even (idealized) automated and perfect detection of abusive content automatically would only solve half the problem, though. What remains is the danger of creating "fake news" that are hard to spot [13]. Regulation arguably needs to tackle these challenges. To better highlight them, for the discussion that follows, we will consider different scenarios of LGAIM use. Consider the following two lead examples: in a business context, one might think of a clothing and sportswear manufacturer (e.g., adidas or Nike) that wants to use the potential of a LGAIM specifically for the design of clothing. For this purpose, adidas might use a pre-trained model provided by a developer (e.g., StabilityAI), while another entity, the deployer, would fine-tune the model according to adidas' requirements (and possibly host it on a cloud platform). As a second exemplary use case, in a private setting, one could think of a young parent that uses an AI text generator to

generate a funny (and suitable) invitation text for her daughter's birthday party. To do so, (s)he might consult ChatGPT and ask the chatbot to come up with an appropriate suggestion

3 DIRECT REGULATION OF THE AI VALUE CHAIN: THE EUROPEAN AI ACT On May 13, 2022, the French Council presidency circulated an amendment to the draft AI Act, Articles 4a-4c, on what the text calls general-purpose AI systems (GPAIS). This novel passage, which did not spark much debate initially, has surreptitiously come to form the nucleus of direct regulation of LGAIMs. It continues to be fiercely contested in the European Parliament [19]. The general approach adopted by the Council on December 6, 2022, now defines GPAIS as systems "intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems" (Article 3(1b) AI Act). To ensure a "fair sharing of responsibilities along the AI value chain" (Recital 12c AI Act), these systems are subjected to the high-risk obligations (e.g., Article 8 to 15 AI Act) if they may be used as high-risk systems or as components thereof (Article 4b(1)(1) and 4b(2) AI Act). These duties arise once the Commission has specified, in implementing acts, how the high-risk rules should be adapted to GPAIS (Article 4b(1) AI Act). The territorial scope of the AI Act extends, inter alia, to providers placing on the market or putting into service AI systems in the EU, and also to situations where the output produced by the system is used in the Union. Hence, these rules may even apply if the provider is entirely based outside of the EU (Article 2(1) AI Act). Exceptionally, the provider is exempted from specific obligations for GPAIS providers if the provider explicitly and publicly excludes all high-risk uses of the GPAIS; however, the exemption fails if the exclusion is not made in good faith (Article 4c(1) and (2) AI Act). Nevertheless, if any provider detects or is informed about market misuse of its system, it must take all proportionate measures to stop the misuse and avoid harm (Article 4c(3) AI Act). This notice and-action mechanism, structurally known from copyright law [50-52] and the DSA [26, 28, 53, 54], complements the novel, active production monitoring obligation (Articles 61, 4(2) AI Act), which had so far only been contained in the product liability tort law of some Member States [55] and will be partially introduced in the product liability upgrade as well (Article 6(1)(c) and (e) Product Liability Directive Proposal[2] [20, 56, 57]).

[2] European Commission, Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products, COM(2022) 495 final.

3.1 Critique of the GPAIS AI Act Rules

The AI Act heroically strives to keep pace with the accelerating dynamics in the AI technology space. However, in our view, the recently introduced rules on GPAIS fail to do justice to the peculiarities of large AI models, and particularly LGAIMs, for three reasons.

3.1.1 Toward a Definition of GPAIS

First, the definition in Article 3(1b) AI Act is significantly over-inclusive. Rules on GPAIS were inspired by the surge in the release of and literature on foundation models and LGAIMs. As seen in Part 2, LGAIMs operate with large numbers of parameters, training data, and compute. While not yet bordering on artificial general intelligence [10], LGAIMs still are more versatile than the narrower deep learning systems that have dominated the third wave of AI so far. Significantly, they can be deployed to solve tasks they have not been specifically trained for [33], and generally operate on a wider range of problems than traditional models. Conceptually, their "generality" may refer to their *ability* (e.g., language

versus vision, or combinations in multimodal models); *domain* of use cases (e.g., educational versus economic); breadth of *tasks* covered (e.g., summarizing versus completing text), or versatility of *output* (e.g., black and white versus multicolored image) [10]. GPAIS, in our view, must necessarily display significant generality in ability, tasks, or outputs, beyond the mere fact that they might be integrated into various use cases (which also holds true for extremely simple algorithms). The broad definition of GPAIS in the AI Act clashes with this understanding, however. According to that rule, every simple image or speech recognition system seems to qualify, irrespective of the breadth of its capabilities; rightly, this only corresponds to a minority position in the GPAIS literature [10, 58].

This problematic overinclusion is caused by the second half-sentence of Article 3(1b) AI Act, where further specifications–use in different contexts and AI systems–are not formulated as (disjunctive) necessary conditions, but as merely possible (and then likely qualifying) examples of GPAIS ("may"). To specifically capture truly general-purpose systems, the definition would have to be revised so that the use of the system in different contexts or for substantively different AI systems are necessary, and not sufficient, conditions. Additionally, it should require that GPAIS display significant generality of ability, task, or output, in decreasing order of relevance [cf. also 10]. As a result, models that display only one set of abilities and tasks would need to have highly diverse output to qualify as GPAIS; conversely, multimodal models would generally qualify, even if they only apply to one specific task and output is not significantly variable.

3.1.2 Risk Management for GPAIS

Second, even such a narrower definition would not avoid other problems. Precisely because large AI models are so versatile, providers will generally not be able to avail themselves of the exception in Article 4c(1) AI Act: by excluding all high-risk uses, they would not act in good faith, as they would have to know that the system, once released, may and likely will be used for at least one high-risk application. For example, language models may be used to summarize or rate medical patient files, student, job, credit or insurance applications (Annexes II, Section A. No. 12, 13 and III No. 3-5 AI Act). Image or video models might be used to visualize safety aspects of high-risk products regulated under the New Legislative Framework (see Annex II Section A. AI Act). Unless any misuse can be verifiably technically excluded, LGAIMs will therefore generally count as high-risk systems. This, however, entails that they have to abide by the high-risk obligations, in particular the establishment of a comprehensive risk management system, according to Article 9 AI Act. Setting up such a system seems to border on the impossible, given LGAIMs' versatility. 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). On this basis, mitigation strategies for all these risks have to be developed and implemented (Article 9(2)(d) and (4) AI Act). 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). 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, 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 exposes LGAIM providers to significant liability risk [20, 56].

3.1.3 Adverse Consequences for Competition

Third, the current GPAIS 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 (https://www.kdnuggets.com/2022/09/john-snow-top-open-source-large-language-models.html). 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, 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 be prohibitively costly. Hence, the AI Act will have the 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 [59]. In this sense, the AI Act threatens to undermine the efforts of the Digital Markets Act[3] to infuse workable competition into the core of the digital and platform economy.

[3] 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).

3.2 Solution and Follow-up Problems: Focus on Deployers and Users

This critique does not imply, of course, that LGAIMs should not be regulated at all. However, in our view, a different approach is warranted. Scholars have noted that the regulatory focus should shift [8, 9] and move towards LGAIM deployers and users, i.e., those calibrating LGAIMs for and using them in concrete high-risk applications.

3.2.1 Terminology: Developers, Deployers, Users, and Recipients

Lilian Edwards has rightly suggested to differentiate between developers of GPAIS, deployers, and end users [8, see also 20]. In the following, we take this beginning differentiation in the AI value chain one step further. In many scenarios, there will be at least four entities involved, in different roles [cf. 60]. We suggest that the terminology in the AI Act and other pertaining regulation must be adapted to the evolving AI value chain in the following way.

- developer: this is the entity originally creating as (pre-) training the model. In the AI Act, this entity is called the provider (Article 3(2)). Real-world examples would be OpenAI, Stability, or Google.
- deployer: this is the entity fine-tuning the model for a specific use case. The AI act calls these entities (professional) user (Article 3(4)) or provider, depending on the circumstances. Note that there could be several deployers (working jointly or consecutively), leading to a true AI value chain similar to OEM value chains. Alternatively, the developer could simultaneously act as a deployer (vertical integration). This would raise the typical competition law issues of vertical integration [61-63].
- user: this is the entity actually generating output from an LGAIM, e.g. via prompts, and putting it to use. The user may harness the output in a professional or a non-professional capacity. Hence, we introduce the following distinction within the category of users:
    - professional user: an entity using AI output for professional purposes. The AI Act calls such entities professional users as well (cf. Art. 2(8) AI Act). The professional user could be a for-profit or a non-profit company, an NGO, an administrative agency, a court, or the legislator, for example. Potential real-world examples would be Nike, Adidas, or any other examples of the types of professional users just listed. Other professional users could be persons like the authors using ChatGPT or other generative of AI systems for academic work and education.
    - non-professional user: an entity using AI output for non-professional purposes. The AI Act calls such entities non-professional users as well (cf. Art. 2(8) AI Act). These could be any person using ChatGPT to brainstorm ideas for her children's birthday party, for example.

  If there was a deployer involved, the user would implement the fine-tuned model, calibrated by the deployer, into its product or use case. Typically, this would be a professional user. In the alternative, the user could harness the LGAIM by directly connecting with the developers. This includes non-professional users (birthday ideas via ChatGPT) as well as professional ones (academic paper writing with the help of ChatGPT; advertisement campaign for company with slogans designed by Aleph Alpha's Luminous).
- recipient: this is the entity consuming the product offered by the professional user. It does not generate LGAIM output. Typically, it will be a consumer, but it could also be a company, an NGO, an administrative agency, the court, or the legislator. However, it sits at the receiving, passive end of the pipeline. In the terminology of the AI Act, these entities would arguably fall in the group of affected persons (see, e.g., Recitals 39 and 42 AI Act). Real-world examples include consumers exposed to AI-generated advertisements and kids/parents invited to AI designed birthday invitations and parties.

Such a shift entails several follow-up problems that need to be addressed to operationalize any focus on deployers [8]. First, deployers and users may be much smaller and less technologically sophisticated than LGAIM developers. This is not a sufficient reason to exempt them from regulation and liability, but it points to the importance of designing a feasible allocation of responsibilities along the AI value chain. Obligations must be structured in such a way that deployers and users can reasonably be expected to comply with them, both by implementing the necessary technological adjustments and by absorbing the compliance costs.

Second, many of the AI Act's high-risk obligations refer to the training and modeling phase conducted, at least partially, by the LGAIM developers. Typically, LGAIM developers will pre-train a large model, which may then be fine-tuned by deployers, potentially in collaboration with developers [64, 65], while users ultimately make the decision what the AI system is used for specifically. To meet the AI Act requirements concerning training data (Article 10),

documentation and record-keeping (Articles 11 and 12), transparency and human oversight (Articles 13 and 14), performance, robustness and cybersecurity (Article 15), and to establish the comprehensive risk management system (Article 9), any person responsible will need to have access to the developer's and deployer's data and expertise. This unveils a regulatory dilemma: focusing on developers entails potentially excessive and inefficient compliance obligations; focusing on deployers and users risks burdening those who cannot comply due to limited insight.

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. In this vein, we suggest a combination of strategies known from pre-trial discovery, trade secrets law, and the GDPR. Under the current 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 [8, 9]. 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 problem of balancing cooperation and disclosure with the protection of information is not limited to the AI Act. It has an internal and external dimension. Internally, i.e., in the relationship between the party requesting and the party granting access, virtually all access rights are countered, by the granting party, by reference to supposedly unsurmountable trade secrets or IP rights [66-68]. 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 [20, 56, 57]. Article 15(4) GDPR contains a similar provision, which by way of analogy also applies to the access right in Article 15(1) GDPR [69, 70]. Extensive literature and practice concerning this problem exists in the realm of the US pretrial discovery system [71-75]. Under this mechanism, partially adopted by the proposed EU evidence disclosure rules [20], 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' 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 NDA's 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 [68, 71, 76] 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)) [76]. With these safeguards in place, LGAIM developers should be compelled, and not merely encouraged, to cooperate with deployers and users if they have authorized the deployment.

Concerning the external dimension, the question arises of who should be responsible for fulfilling pertinent duties and, 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 [8]). 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 disclosure of evidence requests under the AI liability regime. Finally, joint and several liability ensures collaboration and serves the compensation interests of injured persons. 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: 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 the LGAIM developer, deployer, user and the affected persons.

3.3 Non-Discrimination Law

The situation is slightly different with non-discrimination law, which generally applies, in the US as well as the EU, in a technology-neutral way [1, 77-80, 81; but see also Massachusetts Legislature, Bill SD.1827 (193rd) for a (brief) proposal for anti-discrimination rules specifically for LGAIM operators]. Importantly, however, it only covers certain enumerated areas of activity, such as employment, education, or publicly available offers of goods and services [80, 82]. This begs the question whether general-purpose systems may be affected by non-discrimination provisions even before they have been deployed in specific use cases. Concerning EU law, the CJEU has held in a string of judgments (CJEU, Case C-54/07, Feryn; Case C-507/18, Associazione Avvocatura per i diritti LGBTI) that non-discrimination provisions may apply to preparatory activities preceding, e.g., the actual job selection under certain conditions: in the concrete cases, statements made on a radio program indicating an intention not to recruit candidates of a particular sexual orientation are considered "conditions for access to employment" if the relationship between the statement and the employer's recruitment policy is not merely hypothetical (CJEU, Case C-507/18, Associazione Avvocatura per i diritti LGBTI, para. 43). For anti discrimination directives to apply, a preliminary measure must, therefore, concretely relate to an activity covered by the directives.

Regarding the (pre-)training of LGAIMs, the required link arguably exists if the model is specifically prepared for use (also) in discrimination-relevant scenarios (employment; education, publicly available goods or services etc.) [cf. also 83]. Conversely, if a generic LGAIM is developed without any specific link to such scenarios (even though it may be

theoretically used in these cases) non-discrimination law does not apply to the development itself. Again, it obviously applies to the concrete deployment in the respective scenario. If non-discrimination law applies, intricate questions ensue which transcend the scope of this paper, for example concerning concrete proof of discrimination, harm, and standing to sue. Ultimately, we would argue, however, that significant underperformance for legally protected groups will be indicative of (indirect or even direct) discrimination, establishing a prima facie case [1, 77-81].

4 DATA PROTECTION UNDER THE GDPR

A second major challenge for any AI model is GDPR compliance. While the requirements for large generative models

are, arguably, not categorically different from those for any machine learning model, we refrain from a substantiated analysis at this point. However, we note that recent studies have shown that LGAIMs are vulnerable to inversion attacks, even to a greater extent than previous generative models, such as generative adversarial networks (GANs [84]) [23]. As a consequence, data used for training may be reproduced from the model. This is less of a concern for copyrighted material as the new text-and-data-mining exceptions (Articles 3 and 4 Copyright in the Digital Single Market Directive[4]), at least generally and where rightsholders have not opted out, allow the use of publicly accessible copyrighted material for machine learning purposes [85, 86]. However, a legal basis under Article 6 GDPR is necessary for any use of personal data for training [22, 83]. According to some scholars, even the model itself might be considered personal data, considering the possibility of inversion attacks [22]. As consent was typically not obtained [87], developers need to avail themselves of the balancing test and/or the purpose change test (Article 6(1)(f) and (4) GDPR) [88-91], potentially in conjunction with specific research exceptions (Article 89 GDPR) [92-94]. Compliance with the GDPR will depend on a range of factors [95-98], such as the intended purposes of the model, the type of personal data used, the likelihood of model inversion, and the probability of re-identification of concrete data subjects. While scholars are divided about whether the use of personal data for machine learning purposes should be permitted under the balancing test (permissive view: [83]; more restrictive view: [91, 99]; see also [100]), the threat of model inversion arguably weighs in favor of data subjects.

Importantly, if the personal data constitutes sensitive data in the sense of Article 9 GDPR, developers need to invoke, besides Article 6 GDPR, an exception under Article 9(2) GDPR. While there is no universal balancing test available at the EU level, some Member States have introduced (more restrictive) balancing tests for research purposes under the opening clauses of Article 9(2)(g-j) GDPR (see, e.g., § 22 of the German Data Protection Act, BDSG [83]). With respect to sensitive data, model inversion constitutes an even more serious threat to GDPR compliance as the risk of a reproduction of sensitive data will hardly be overcome by legitimate interests of developers, unless the model serves truly critical purposes (potentially in medicine or emergency cases).

5 GENERATIVE MODEL CONTENT MODERATION: THE EUROPEAN DATA SERVICES ACT The third large regulatory frontier concerning LGAIMs is content moderation. Generative models, as virtually any novel technology, may be used for better or worse purposes [101]. The developers of ChatGPT, specifically, anticipated the potential for abuse and trained in internal AI moderator, with controversial help from Kenyan contractors [49] , to detect and block harmful content [12]. AI Research has made progress in this area recently [102-104]. OpenAI has released a content filtering mechanism which users may apply to analyze and flag potentially problematic content along several categories (violence; hate; sexual content etc.).[5] Other large generative models have similar functionalities. However,

[4] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market, OJ L 130, 17.5.2019, 92.
[5] See https://platform.openai.com/docs/api-reference/moderations.

actors intent on using ChatGPT, and other models, to generate fake or harmful content will find ways to prompt them to do just that. Prompt engineering is becoming a new art to elicit any content from LGAIMs [105] and fake news is harder to detect than hate speech, even though industry efforts are underway via increased model and source transparency [106]. As could be expected, DIY instructions for circumventing content filters are already populating YouTube and reddit,[6] and researchers have already generated an entire hate-filled shitstorm, along with code for proliferation, using ChatGPT [5]. The propensity of ChatGPT particularly to hallucinate when it does not find ready-made answers can be exploited to generate text devoid of any connection to reality, but written in the style of utter confidence, persuasion, scholarly

attitude, or derision, as conditions warrant. While the European Parliament is investigating LENSA for inappropriate content generation [107, 108] the timing of the advent of truly powerful LGAIMs could hardly be any more favorable for malicious actors. The Russian attack on Ukraine, the Corona pandemic, climate change, and the political feuds in the US and beyond already fuel hate crimes and fake news [109, 110]. This toxic political climate now meets the factual demolition of content moderation on Twitter under the auspices of its new owner. LGAIMs could well be a powerful instrument in the upcoming election cycles to target individual actors and sway public opinion. They allow for the automated mass production, and proliferation, of highly sophisticated, seemingly fact-based, but actually utterly nonsensical fake news and harmful speech campaigns.

To stem the tide of such phenomena, the EU has recently enacted the DSA. However, it was designed to mitigate illegal content on social networks, built by human actors or the occasional Twitter bot, not to counter LGAIMs. The problem lies not in its territorial applicability: the DSA, like the AI Act, covers services offered to users in the EU, irrespective of where the providers have their place of establishment (Article 2(1), 3 (d) and (e) DSA). Platforms like Facebook or Twitter must implement a notice and action system where users can report potentially illegal content that will then be reviewed and removed if found illegal (Article 16 DSA). Larger platforms must have an internal complaint and redress system (Article 20 DSA) and provide out-of-court dispute resolution (Article 21 DSA). Hence, users are spared the lengthy process of going to court to challenge problematic content. Repeat offenders of content moderation policies risk having their accounts suspended (Article 23 DSA). Content highlighted as problematic by so-called trusted flaggers, which have to register with Member States, must be prioritized and decided upon without undue delay (Article 22 DSA). Very large online platforms (> 45 million active users) must also implement a comprehensive compliance system, including proactive risk management and independent audits (Articles 33-37 DSA).

Passed with the best of intentions, the DSA, however, seems outdated at the moment of its enactment. This results from two crucial limitations in its scope of application. First, it covers only so-called intermediary services (Article 2(1) and (2) DSA). Article 3(g) DSA defines them as “mere conduit” (e.g., Internet access providers), “caching” or “hosting” services (e.g., social media platforms, see also Recital 28 DSA). Arguably, however, LGAIMs do not match any of these categories. Clearly, they are not comparable to access or caching service providers, which power Internet connections. Hosting services, in turn, are defined as providers storing information provided by, and at the request of, a user (Article 3(g)(iii) DSA) [see also 111].While users do request information from LGAIMs via prompts, they can hardly be said to provide this information. Rather, other than in traditional social media constellations, it is the LGAIM, not the user, who produces the text. To the contrary, CJEU jurisprudence shows that even platforms merely storing user-generated content may easily lose their status as hosting providers, and concomitant liability privileges under the DSA (and its predecessor in this respect, the E-Commerce Directive), if they “provide assistance” and thus leave their “neutral position”, which may even mean

[6] See, e.g., https://www.youtube.com/watch?v=qpKlnYLtPjc; https://www.reddit.com/r/OpenAI/comments/zjyrvw/a_tutorial_on_how_to_use_chatgpt_to_make_any/.

merely promoting user-generated content (CJEU, Case C-324/09, L'Oréal para 116). A fortiori, systems generating the content themselves cannot reasonably be qualified as hosting service providers. Hence, the DSA does not apply. This does not imply that LGAIM content generation is not covered by content liability laws. Rather, its output may be covered by speech regulation, similar to comments made by human users online. However, this branch of the law is largely left to Member State tort law, with the exception of Article 82 GDPR in the case of processing personal data of victims, which seems rather far-fetched in LGAIM constellations. Not only does such direct speech regulation vary considerably

between Member States [112]: one Member State (Germany) even declined to enforce a judgment of another Member States' court (Poland) on speech regulation, as it saw a collision with its constitutional provisions on free speech, amounting to a breach of its own *ordre public* (German Constitutional Court, Judgment of 19 July 2018, Case IX ZB 10/18). It also often lacks precisely the instruments the DSA has introduced to facilitate the rapid yet procedurally adequate removal of harmful speech and fake news from the online world: notice and action mechanisms flanked by procedural safeguards; trusted flaggers; obligatory dispute resolution; and comprehensive compliance and risk management regimes for large platforms.

The risk of a regulatory loophole might be partially closed, one might object, by the applicability of the DSA to LGAIM generated posts that human users, or bots, publish on social networks. Here, the DSA generally applies, as Twitter et al. qualify as hosting service providers. However, a second important gap looms: Recital 14 DSA specifies that the main part of the regulation does not cover "private messaging services." While the notice and action mechanism applies to all hosting services, instruments like trusted flaggers, obligatory dispute resolution, and risk management systems are reserved for the narrower group of "online platforms" [113]. To qualify, these entities must disseminate information to the public (Article 3(g)(iii), (k) DSA). According to Recital 14 DSA, closed groups on WhatsApp and Telegram, on which problematic content particularly proliferates, are explicitly excluded from the DSA's online platform regulation (Articles 19 ff. DSA) as messages are not distributed to the general public. With the right lines of codes, potentially supplied by an LGAIM as well [5], malicious actors posting content in such groups may therefore fully escape the ambit, and the enforcement tools, of the DSA.

Hence, the only action to which the full range of the DSA mechanisms continues to apply is the posting of LGAIM generated content on traditional social networks, such as Twitter, YouTube, or Instagram. However, at this point in time, Pandora's box has already been opened. Misinformation may also be spread effectively and widely via interpersonal communication. Even if the EU legislator has decided to exclude closed groups from the scope of the DSA [114], this balance needs to be reassessed in the context of readily available LGAIM output, which exacerbates risks. Even the most stringent application of DSA enforcement mechanisms, potentially coupled with GDPR provisions on erasure of data (Article 17(2) and 19 GDPR), cannot undo the harm done, and often cannot prevent the forward replication of problematic content [115]. Overall, current EU law, despite the laudable efforts in the DSA to mitigate the proliferation of fake news and hate speech, fails to adequately address the dark side of LGAIMs.

6 POLICY PROPOSALS

The preceding discussion has shown that regulation 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 feasible for LGAIM developers large and small to avoid a winner-takes-all scenario and further market concentration [59]. This is crucial not only for innovation and or consumer welfare [28, 116, 117], but also for environmental sustainability. While the carbon footprint of IT and AI is significant and steadily rising [38-42], and training

of LGAIMs is particularly resource intensive [118], large models may ultimately create significantly fewer greenhouse gas emissions than their smaller brethren [118, p. 15].

Against this background, we make four concrete, workable suggestions for LGAIM regulation: (i) transparency obligations; (ii) mandatory yet limited risk management; (iii) non-discrimination data audits; and (iv) expanded content moderation.

6.1 Transparency

The AI Act contains a wide range of disclosure obligations (Article 11, Annex IV AI Act) that apply, however, only to high-risk systems. In our view, given the vast potential and growing relevance of LGAIMs for many sectors of society, LGAIMs should — irrespective of their categorization as high-risk or non-high-risk — be subject to two distinct transparency duties. First, *LGAIM developers and deployers* should be required to report on their performance metrics as well as any incidents and mitigation strategies concerning harmful content. Ideally, to the extent technically feasible [38, 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 [20, p. 65 f., see also 119].

Second, *professional users* should be obligated to disclose which parts of their publicly available content were generated by or adapted from LGAIMs. 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. 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 [120]. The European Parliament is currently pondering a watermark obligation for generative AI [119]. Importantly, more interdisciplinary research is necessary to develop markings that are easy to use and recognize, but hard to remove by average users [121]. This should be coupled

with research on AI-content detection to highlight such output where watermarks fail [103, 122].
(https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text/)

6.2 Risk Management and Staged Release

As mentioned, one major obstacle to the effective application of the AI Act to LGAIMs proper is comprehensive risk

management. Here, novel approaches are needed. Scholars have rightly suggested that powerful models should be released consciously, trading off the added benefit of public scrutiny with the added risk of misuse in the case of full public releases [123]. In our view, a limited, staged release, coupled with only access for security researchers and selected stakeholders, will often be preferable [see also 6, 123-125]. This adds a nuanced, community-based risk management strategy by way of codes of conduct to the regulatory mix [cf. also 125]. Regulatory oversight could be added by way of "regulated self regulation;" an approach with potentially binding effect of the code of conduct, à la Article 40 GDPR, seems preferable to the purely voluntary strategy envisioned in Article 69 AI Act.

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 (see Part 3.2). This strategy aligns with a general principle of product safety law [9]: 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[7]–but not if they are sold in the local DIY store for generic use. The same principle should be applied to LGAIMs.

6.3 Non-Discrimination Data Audits

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. Here, it seems paramount to mitigate the risk at its roots. 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.

6.4 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. 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

[7] See, e.g., product standards, aerospace series, DIN EN 4845–4851 (December 2022) on screws.

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.

Like in the other areas, more fundamental research is additionally needed to not only mark AI-generated content, but to integrate adherence to facts and content moderation into the models themselves. Transparent LGAIMs connected to (good-old) knowledge bases, such as AtMan [43], present a promising way forward in this direction.

Overall, we have added several policy proposals. As a matter of regulatory technique, the legislator should, in our view, strive to shift its strategy from technology-specific regulation–which will often be outdated before eventually enacted–toward more technology-neutral regulation wherever possible. As seen, non-discrimination law, formulated in a technology-neutral way, continues to grapple with various challenges, but arguably does a better job capturing the dynamics of LGAIM development than the AI Act or the DSA, at least in the way they are currently enacted and proposed.

7 CONCLUSION

Scholars and regulators have long suggested that technology-neutral laws may be better prepared to tackle emerging risks given the rapid pace of innovation in machine learning [126-128]. While this claim, arguably, cannot be generally affirmed or refuted, LGAIMs offer a cautionary example for regulation focused specifically on certain technologies. As our study shows, technology-neutral laws sometimes fare better because technology-specific regulation (on platforms; AI systems) may be outdated before (AI Act, AI liability regime) or at the moment of its enactment (DSA). Overall, we add several policy proposals to the emerging regulatory landscape surrounding LGAIMs.

First, rules in the AI Act and other direct regulation must match the specificities of pre-trained models. This includes generally singling out concrete high-risk applications, and not the pre-trained model itself, as the object of high-risk obligations. For example, it seems inefficient and practically infeasible to compel the developers of ChatGPT to draw up a comprehensive risk management system covering, and mitigating, all risks to health, safety and fundamental rights ChatGPT may pose. Rather, if used for a concrete high-risk purpose (e.g., summarizing or grading résumés in employment decisions), the specific deployer should have to comply with the AI Act's high-risk obligations, including the risk management system. The devil, however, is in the detail: providers need to cooperate with deployers to comply with even such narrower regulatory requirements. Here, we suggest drawing on experience from the US pretrial discovery system to balance interests in the access to information with trade secret protection.

Second, exceptionally, non-discrimination provisions, including a version of Article 10 AI Act, should apply to LGAIM developers. In this way, biased output can arguably be prevented most effectively. This particularly concerns the collection and curation of training data scraped from the Internet.

Third, detailed transparency obligations are warranted. This concerns both LGAIM developers (performance metrics; harmful speech issues arisen during pre-training) and users (disclosure of the use of LGAIM-generated content). Fourth, the core of the DSA's content moderation rules should be expanded to cover LGAIMs. This includes notice and action mechanisms, trusted flaggers, and comprehensive audits. Arguably, it is insufficient to tackle AI-generated hate speech and fake news ex post, once they are posted to social media. At this point, their effect will be difficult to stop. Rather, AI generation itself must be moderated by an adequate combination of AI tools, developer and user interventions, and law.

In all areas, regulators and lawmakers need to act fast to keep track with the unchained dynamics of ChatGPT et al. Updating regulation is necessary both to maintain the civility of online discourses and to create a level playing field for developing and deploying the next generation of AI models, in the EU and beyond.

ACKNOWLEDGMENTS

Passages taken over from ChatGPT are found in the section on technical foundations, all italicized and marked with "",

and referenced by the prompt used. They were all collected on January 17, 2023. We deem them factually correct unless otherwise noted. This paper benefitted from comments by Johannes Otterbach and audiences at Bucerius Law School (Hamburg) and the University of Hamburg. All errors remain entirely our own.

8 REFERENCES

[1] Zuiderveen Borgesius, F. J. Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights*, 24, 10 (2020), 1572-1593.
[2] Lee, D. and Yoon, S. N. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *International Journal of Environmental Research and Public Health*, 18, 1 (2021), 271. [3] Aung, Y. Y., Wong, D. C. and Ting, D. S. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *British medical bulletin*, 139, 1 (2021), 4-15. [4] Marcus, G. *A Skeptical Take on the A.I. Revolution*. The Ezra Klein Show, The New York Times, City, 2023. [5] Beuth, P. *Wie sich ChatGPT mit Worten hacken lässt*. Der Spiegel, City, 2023.
[6] Bergman, A. S., Abercrombie, G., Spruit, S., Hovy, D., Dinan, E., Boureau, Y.-L. and Rieser, V. *Guiding the release of safer E2E conversational AI through value sensitive design*. Association for Computational Linguistics, City, 2022. [7] Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W. and Elovici, Y. The threat of offensive ai to organizations. *Computers & Security* (2022), 103006.
[8] Edwards, L. Regulating AI in Europe: four problems and four solutions. *Retrieved March*, 15 (2022), 2022. [9] Hacker, P., Engel, A. and List, T. *Understanding and regulating ChatGPT, and other large generative AI models*. City, 2023.
[10] Gutierrez, C. I., Aguirre, A., Uuk, R., Boine, C. C. and Franklin, M. A Proposal for a Definition of General Purpose Artificial Intelligence Systems. *Working Paper, https://ssrn.com/abstract=4238951* (2022). [11] Heikkilä, M. *The EU wants to regulate your favorite AI tools*. City, 2023.
[12] KI-Bundesverband *Large European AI Models (LEAM) as Leuchtturmprojekt für Europa*. City, 2023. [13] Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M. and Sedova, K. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv preprint arXiv:2301.04246* (2023).
[14] Chee, F. Y. and Mukherjee, S. *Exclusive: ChatGPT in spotlight as EU's Breton bats for tougher AI rules*. Reuters, City, 2023.
[15] Smith, B. *Meeting the AI moment: advancing the future through responsible AI*. City, 2023.
[16] Lieu, T. *I'm a Congressman Who Codes. A.I. Freaks Me Out.*, City, 2023.
[17] *An Act drafted with the help of ChatGPT to regulate generative artificial intelligence models like ChatGPT.*, City, 2023.
[18] Veale, M. and Borgesius, F. Z. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22, 4 (2021), 97-112. [19] Bertuzzi, L. *Leading MEPs exclude general-purpose AI from high-risk categories – for now*. City, 2022. [20] Hacker, P. The European AI Liability Directives - Critique of a Half-Hearted Approach and Lessons for the Future. *Working Paper, https://arxiv.org/abs/2211.13960* (2022).
[21] Fredrikson, M., Jha, S. and Ristenpart, T. *Model inversion attacks that exploit confidence information and basic countermeasures*. City, 2015.
[22] Veale, M., Binns, R. and Edwards, L. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376, 2133 (2018), 20180083.

[23] Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D. and Wallace, E. Extracting Training Data from Diffusion Models. *arXiv preprint arXiv:2301.13188* (2023).
[24] Douek, E. Content Moderation as Systems Thinking. *Harv. L. Rev.*, 136 (2022), 526.
[25] De Gregorio, G. Democratising online content moderation: A constitutional framework. *Computer Law & Security*

Review, 36 (2020), 105374.
[26] Heldt, A. P. *EU Digital Services Act: The white hope of intermediary regulation*. Palgrave, City, 2022.
[27] Meyer, P. *ChatGPT: How Does It Work Internally?* , City, 2022.
[28] Eifert, M., Metzger, A., Schweitzer, H. and Wagner, G. Taming the giants: The DMA/DSA package. *Common Market Law Review*, 58, 4 (2021), 987-1028.
[29] Laux, J., Wachter, S. and Mittelstadt, B. Taming the few: Platform regulation, independent audits, and the risks of capture created by the DMA and DSA. *Computer Law & Security Review*, 43 (2021), 105613. [30] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A. and Brunskill, E. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021). [31] Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D. and Elhage, N. Predictability and surprise in large generative models. *ACM Conference on Fairness, Accountability, and Transparency* (2022), 1747-1764.
[32] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J. and Clark, A. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* (2022). [33] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. and Askell, A. Language models are few-shot learners. *Advances in neural information processing systems*, 33 (2020), 1877-1901.
[34] Kim, B., Kim, H., Lee, S.-W., Lee, G., Kwak, D., Jeon, D. H., Park, S., Kim, S., Kim, S. and Seo, D. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. *arXiv preprint arXiv:2109.04650* (2021).
[35] Bienert, J. and Klös, H.-P. *Große KI-Modelle als Basis für Forschung und wirtschaftliche Entwicklung*. IW Kurzbericht, 2022.
[36] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the Proceedings of the 38th International Conference on Machine Learning* (Proceedings of Machine Learning Research, 2021). PMLR, [insert City of Publication],[insert 2021 of Publication].
[37] Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T. and Wu, Y. Combined scaling for open-vocabulary image classification. *arXiv e-prints* (2021), arXiv: 2111.10050. [38] OECD *Measuring the Environmental Impacts of AI Compute and Applications: The AI Footprint*. City, 2022. [39] Freitag, C., Berners-Lee, M., Widdicks, K., Knowles, B., Blair, G. S. and Friday, A. The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns*, 2, 9 (2021), 100340. [40] ACM, T. P. C. *ACM TechBrief: Computing and Climate Change*. City, 2021.
[41] Cowls, J., Tsamados, A., Taddeo, M. and Floridi, L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. *AI & Society* (2021), 1-25.
[42] Taddeo, M., Tsamados, A., Cowls, J. and Floridi, L. Artificial intelligence and the climate emergency: Opportunities, challenges, and recommendations. *One Earth*, 4, 6 (2021), 776-779.
[43] Ananthaswamy, A. *The Physics Principle That Inspired Modern AI Art*. City, 2023.
[44] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S. *Deep unsupervised learning using nonequilibrium thermodynamics*. PMLR, City, 2015.
[45] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. and Neubig, G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55 (2021), 1 - 35. [46] Luccioni, A. S. and Viviano, J. D. What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus. *arXiv preprint arXiv:2105.02732* (2021).
[47] Nadeem, M., Bethke, A. and Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).
[48] Zhao, Z., Wallace, E., Feng, S., Klein, D. and Singh, S. *Calibrate before use: Improving few-shot performance of language models*. PMLR, City, 2021.
[49] Perrigo, B. *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic*. City, 2023.

[50] Bar-Ziv, S. and Elkin-Koren, N. Behind the scenes of online copyright enforcement: Empirical evidence on notice &

takedown. *Conn. L. Rev.*, 50 (2018), 339.
[51] Cobia, J. The digital millennium copyright act takedown notice procedure: Misuses, abuses, and shortcomings of the process. *Minn. JL Sci. & Tech.*, 10 (2008), 387.
[52] Urban, J. M. and Quilter, L. Efficient process or chilling effects-takedown notices under Section 512 of the Digital Millennium Copyright Act. *Santa Clara Computer & High Tech. LJ*, 22 (2005), 621.
[53] Hacker, P., Cordes, J. and Rochon, J. Regulating Gatekeeper AI and Data: Transparency, Access, and Fairness under the DMA, the GDPR, and beyond. *Working Paper, https://arxiv.org/abs/2212.04997* (2022). [54] Morais Carvalho, J., Arga e Lima, F. and Farinha, M. Introduction to the Digital Services Act, Content Moderation and Consumer Protection. *Revista de Direito e Tecnologia*, 3, 1 (2021), 71-104.
[55] Riehm, T. and Meier, S. Product Liability in Germany. *J. Eur. Consumer & Mkt. L.*, 8 (2019), 161. [56] Spindler, G. Die Vorschläge der EU-Kommission zu einer neuen Produkthaftung und zur Haftung von Herstellern und Betreibern Künstlicher Intelligenz. *Computer und Recht* (2022), 689-704.
[57] Wagner, G. Liability Rules for the Digital Age - Aiming for the Brussels Effect. *European Journal of Tort Law (forthcoming)* (2023), , https://ssrn.com/abstract=4320285.
[58] Bennett, C. C. and Hauser, K. Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. *Artificial intelligence in medicine*, 57, 1 (2013), 9-19.
[59] Geradin, D., Karanikioti, T. and Katsifis, D. GDPR Myopia: how a well-intended regulation ended up favouring large online platforms. *European Competition Journal*, 17, 1 (2021), 47-92.
[60] Bornstein, M., Appenzeller, G. and Casado, M. *Who Owns the Generative AI Platform?* , City, 2023. [61] Blair, R. D. and Kaserman, D. L. *Law and economics of vertical integration and control*. Academic Press, 2014. [62] Perry, M. K. Vertical integration: Determinants and effects. *Handbook of industrial organization*, 1 (1989), 183-255. [63] Kolasky, W. J. and Dick, A. R. The merger guidelines and the integration of efficiencies into antitrust review of horizontal mergers. *Antitrust LJ*, 71 (2003), 207.
[64] Blaschke, T. and Bajorath, J. Fine-tuning of a generative neural network for designing multi-target compounds. *Journal of Computer-Aided Molecular Design*, 36, 5 (2022/05/01 2022), 363-371.
[65] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P. and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).
[66] Drexl, J., Hilty, R., Desaunettes-Barbero, L., Globocnik, J., Gonzalez Otero, B., Hoffmann, J., Kim, D., Kulhari, S., Richter, H. and Scheuerer, S. Artificial Intelligence and Intellectual Property Law-Position Statement of the Max Planck Institute for Innovation and Competition of 9 April 2021 on the Current Debate. *Max Planck Institute for Innovation & Competition Research Paper*, 21-10 (2021).
[67] Calvin, N. and Leung, J. Who owns artificial intelligence? A preliminary analysis of corporate intellectual property strategies and why they matter. *Future of Humanity Institute, February* (2020).
[68] Deeks, A. The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119, 7 (2019), 1829-1850.
[69] Hacker, P. and Passoth, J.-H. Varieties of AI Explanations under the Law. From the GDPR to the AIA, and Beyond. *International Conference on Extending Explainable AI Beyond Deep Models and Classifiers* (2022), 343-373. [70] Schmidt-Wudy, F. *Art. 15 DSGVO*. City, 2023.
[71] McKown, J. R. Discovery of Trade Secrets. *Santa Clara Computer & High Tech. LJ*, 10 (1994), 35. [72] Roberts, J. Too little, too late: Ineffective assistance of counsel, the duty to investigate, and pretrial discovery in criminal cases. *Fordham Urb. LJ*, 31 (2003), 1097.
[73] Shepherd, G. B. An empirical study of the economics of pretrial discovery. *International Review of Law and Economics*, 19, 2 (1999), 245-263.
[74] Subrin, S. N. Discovery in Global Perspective: Are We Nuts. *DePaul L. Rev.*, 52 (2002), 299. [75] Kötz, H. Civil justice systems in Europe and the United States. *Duke J. Comp. & Int'l L.*, 13 (2003), 61. [76] Daniel, P. F. Protecting Trade Secrets from Discovery. *Tort & Ins. LJ*, 30 (1994), 1033. [77] Adams‐Prassl, J., Binns, R. and Kelly‐Lyth, A. Directly Discriminatory Algorithms. *The Modern Law Review* (2022). [78] Wachter, S. The Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law. *arXiv preprint arXiv:2205.01166* (2022).
[79] Wachter, S., Mittelstadt, B. and Russell, C. Why fairness cannot be automated: Bridging the gap between EU non discrimination law and AI. *Computer Law & Security Review*, 41 (2021), 105567.
[80] Hacker, P. Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, 55, 4 (2018), 1143-1186.

[81] Barocas, S. and Selbst, A. D. Big data's disparate impact. *California law review* (2016), 671-732. [82] Wachter, S. Affinity profiling and discrimination by association in online behavioral advertising. *Berkeley Tech. LJ*, 35 (2020), 367.
[83] Hacker, P. A legal framework for AI training data—from first principles to the Artificial Intelligence Act. *Law, Innovation and Technology*, 13, 2 (2021), 257-301.
[84] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63, 11 (2020), 139-144.
[85] Geiger, C., Frosio, G. and Bulayenko, O. The exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market-legal aspects. *Centre for International Intellectual Property Studies (CEIPI) Research Paper*, 2018-02 (2018).
[86] Rosati, E. The exception for text and data mining (TDM) in the proposed Directive on Copyright in the Digital Single Market: technical aspects. *European Parliament* (2018).
[87] Mourby, M., Cathaoir, K. Ó. and Collin, C. B. Transparency of machine-learning in healthcare: The GDPR & European health law. *Computer Law & Security Review*, 43 (2021), 105611.
[88] Zuiderveen Borgesius, F. J., Kruikemeier, S., Boerman, S. C. and Helberger, N. Tracking walls, take-it-or-leave-it choices, the GDPR, and the ePrivacy regulation. *Eur. Data Prot. L. Rev.*, 3 (2017), 353.
[89] Gruschka, N., Mavroeidis, V., Vishi, K. and Jensen, M. *Privacy issues and data protection in big data: a case study analysis under GDPR*. IEEE, City, 2018.
[90] Forgó, N., Hänold, S. and Schütze, B. The principle of purpose limitation and big data. *New technology, big data and the law* (2017), 17-42.
[91] Zarsky, T. Z. Incompatible: The GDPR in the age of big data. *Seton Hall L. Rev.*, 47 (2016), 995. [92] Mondschein, C. F. and Monda, C. The EU's General Data Protection Regulation (GDPR) in a research context. *Fundamentals of clinical data science* (2019), 55-71.
[93] Peloquin, D., DiMaio, M., Bierer, B. and Barnes, M. Disruptive and avoidable: GDPR challenges to secondary research uses of data. *European Journal of Human Genetics*, 28, 6 (2020), 697-705.
[94] Staunton, C., Slokenberga, S. and Mascalzoni, D. The GDPR and the research exemption: considerations on the necessary safeguards for research biobanks. *European Journal of Human Genetics*, 27, 8 (2019), 1159-1167. [95] Gil González, E. and De Hert, P. *Understanding the legal provisions that allow processing and profiling of personal data—an analysis of GDPR provisions and principles*. Springer, City, 2019.
[96] Donnelly, M. and McDonagh, M. Health research, consent and the GDPR exemption. *European journal of health law*, 26, 2 (2019), 97-119.
[97] Jones, M. L. and Kaminski, M. E. An American's Guide to the GDPR. *Denv. L. Rev.*, 98 (2020), 93. [98] Hildebrandt, M. *Law for computer scientists and other folk*. Oxford University Press, 2020. [99] Bonatti, P. A. and Kirrane, S. *Big Data and Analytics in the Age of the GDPR*. IEEE, City, 2019. [100] Butterworth, M. The ICO and artificial intelligence: The role of fairness in the GDPR framework. *Computer Law & Security Review*, 34, 2 (2018), 257-268.
[101] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T. and Filar, B. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228* (2018).
[102] Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C. A., Bull, P., Lipstein, G., Nelli, T. and Zhu, R. *The hateful memes challenge: Competition report*. PMLR, City, 2021.
[103] Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P. and Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33 (2020), 2611-2624.
[104] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F. and Choi, Y. Defending against neural fake news. *Advances in neural information processing systems*, 32 (2019).
[105] Seeha, S. *Prompt Engineering and Zero-Shot/Few-Shot Learning [Guide]*. City, 2022. [106] Deb, M., Deiseroth, B., Weinbach, S., Schramowski, P. and Kersting, K. AtMan: Understanding Transformer Predictions Through Memory Efficient Attention Manipulation. *arXiv preprint arXiv:2301.08110* (2023). [107] Heikkilä, M. *The viral AI avatar app Lensa undressed me—without my consent*. City, 2022. [108] Bertuzzi, L. and Killeen, M. *Tech Brief: Meta's advertising business, next six month's agenda*. City, 2023. [109] OECD. *Disinformation and Russia's war of aggression against Ukraine: Threats and governance responses*. 2022.

[110] Balakrishnan, V., Zhen, N. W., Chong, S. M., Han, G. J. and Lee, T. J. Infodemic and fake news–A comprehensive overview of its global magnitude during the COVID-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction* (2022), 103144.
[111] European, C., Directorate-General for Communications Networks, C., Technology, Hoboken, J., Quintais, J., Poort, J. and Eijk, N. *Hosting intermediary services and illegal content online : an analysis of the scope of article 14 ECD in light of developments in the online service landscape : final report*. Publications Office, 2019. [112] Brüggemeier, G., Ciacchi, A. C. and O'Callaghan, P. *Personality rights in european tort law*. cambridge university press, 2010.
[113] Wilman, F. The Digital Services Act (DSA)-An Ooverview. *Available at SSRN 4304586* (2022). [114] Gerdemann, S. and Spindler, G. Das Gesetz über digitale Dienste (Digital Services Act) (Part 2). *Gewerblicher Rechtsschutz und Urheberrecht* (2023), 115-125.
[115] Korenhof, P. and Koops, B.-J. Gender Identity and Privacy: Could a Right to Be Forgotten Help Andrew Agnes Online? *Working Paper, https://ssrn.com/abstract=2304190* (2014).
[116] Lianos, I. and Motchenkova, E. Market dominance and search quality in the search engine market. *Journal of Competition Law & Economics*, 9, 2 (2013), 419-455.
[117] Geroski, P. A. and Pomroy, R. Innovation and the evolution of market structure. *The journal of industrial economics* (1990), 299-314.
[118] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M. and Dean, J. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350* (2021).
[119] Bertuzzi, L. *AI Act: MEPs want fundamental rights assessments, obligations for high-risk users*. City, 2023. [120] Grinbaum, A. and Adomaitis, L. The Ethical Need for Watermarks in Machine-Generated Language. *arXiv preprint arXiv:2209.03118* (2022).
[121] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I. and Goldstein, T. A Watermark for Large Language Models. *arXiv preprint arXiv:2301.10226* (2023).
[122] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D. and Finn, C. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *arXiv preprint arXiv:2301.11305* (2023).
[123] Liang, P., Bommasani, R., Creel, K. and Reich, R. *The time is now to develop community norms for the release of foundation models*. City, 2022.
[124] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W. and Kreps, S. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203* (2019).
[125] Crootof, R. Artificial intelligence research needs responsible publication norms. *Lawfare Blog* (2019). [126] Hoffmann-Riem, W. *Innovation und Recht-Recht und Innovation: Recht im Ensemble seiner Kontexte*. Mohr Siebeck, 2016.
[127] Bennett Moses, L. Regulating in the face of sociotechnical change (2016).
[128] Bennett Moses, L. Recurring dilemmas: The law's race to keep up with technological change. *U. Ill. JL Tech. & Pol'y* (2007), 239.

APPENDIX H1: PROMPTS

Prompt 1: What are large generative AI models?

Prompt 2: What distinguishes large generative AI models from other AI systems?

Prompt 3: Can you explain the technical foundations of large generative models in simple terms, so that an inexperienced reader understands it?

Prompt 4: What are the objectives, what are the obstacles when it comes to content moderation within large generative AI models?

Prompt 5: How does content moderation work at ChatGPT?