Understanding and Regulating ChatGPT, and Other Large Generative AI Models

Philipp Hacker 2023-01-20T14:20:21

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. This post benefitted from strictly human comments by Jeremias Adams-Prassl and Ulrike Klinger. All errors remain entirely our own.

Large generative AI models (LGAIMs) are shaking up the research community, and society at large, rapidly changing the way we communicate, illustrate, and create. New models are unveiled on almost a daily basis (see, e.g., Google's recent Muse), pushing the state-of-the-art to ever new frontiers, while incumbents are harnessed by millions of users to generate human-level text (e.g., ChatGPT), images (e.g., DALL·E 2), videos (e.g., Synthesia), and even art (e.g., Stable Diffusion).

What has rarely been noticed, however, is that the EU, since the spring of 2022, has quietly been preparing far-reaching rules to explicitly regulate LGAIMs, even if their providers are based outside of the EU. First and foremost, this concerns rules on so-called General-Purpose AI Systems (GPAIS) in the proposed EU AI Act, which are currently hotly debated both in the European Parliament and in the tech community. Significantly, LGAIMs, such as ChatGPT, may take manipulation, fake news, and hate speech to unprecedented levels if not properly reined in, marking a crucial new frontier for trustworthy AI. While the EU has recently enacted the Digital Services Act (DSA), it was passed with human users, or your favorite Twitter bot, posting on social networks in mind, not large AI models running wild.

This post therefore examines the emerging regulatory landscape surrounding large AI models, and makes suggestions for a legal framework that serves the AI community, the users, and society at large. To do so, it proceeds in three major steps. First, we sketch some technical foundations of state-of-the-art large AI models to the extent that this is necessary for the following discussion. The second part delves into the developing regulation, reviewing and critiquing the proposed EU AI Act and the already enacted DSA. The third and final part then makes three concrete proposals on how to better regulate large AI models with a view to (i) transparency obligations (including for sustainability); (ii) mandatory yet limited risk management; and (iii) expanded content moderation.

Technical Foundations of Large Generative AI Models

LGAIMs "*are advanced machine learning models that are trained to generate new data, such as text, images, or audio".*[1)] *This "makes them distinct from other AI models* [… *only*] *designed to make predictions or classifications*"[2)] and is one of the reasons why the size of their training datasets significantly exceeds the amount of training data needed for other AI models. LGAIMs are then trained using various techniques "*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*"[3)] – and novel content mixing and repurposing training data examples in striking, outlandish, or truly surprising ways. The technical intuition behind this is to represent training data (for example, images) as probability distributions, and then to blend distributions of different data types (for example Van Gogh and Rembrandt paintings) to create new output. Large Language Models, like GPT-3, are added to enable output generation based on human text input (prompts, such as: "write an essay about successful entrepreneurs"). However, as the output is ultimately based on training data often found openly on the Internet, LGAIM-generated content may reflect societal biases and prejudice if not properly curated.

Hence, a major challenge for the developers of LGAIMs is the implementation of effective content moderation mechanisms to prevent the generation of harmful or inappropriate content. This is because "*the 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.*"[4)] As if this wasn't enough, "*the sheer volume of content generated by these models can make it difficult to manually review and moderate all of the generated content.*"[5)]

With these problems in mind, the creators of ChatGPT have used "*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.*"[6)] However, even if the creation of harmful or inappropriate content could thus be successfully prevented, the generation of fake news induced by user prompts arguably cannot. This raises important regulatory challenges.

Regulating Large Generative Models: The European AI Act

In May 2022, the French Council presidency made an amendment to the compromise text of the AI Act discussed in the Council of the EU. At that time, it was not much debated outside of specialist circles. The amendment concerned GPAIS, defined as models able to perform well on a broad variety of tasks they had not been explicitly trained for. Importantly, the definition covers the new generation of LGAIMs. This seemingly innocuous amendment has thus not only become the nucleus of LGAIM regulation, but also one of the most contested provisions of the AI Act altogether. Shortly put, the version of the AI Act agreed[7)] upon by the Council stipulates that any GPAIS which may be used for *any* high-risk application (such as employment; medicine; credit scoring; administration; law enforcement; see Article 6 and Annexes II and III AI Act) must, *prima facie*, comply with the full load of the AI Act's obligations for high-risk systems (Article 4b AI Act). This not only concerns performance testing for all intended purposes (Article 15(1) AI Act), but also the establishment of an encompassing risk management system (Article 9 AI Act). These rules apply irrespective of the location of the AI developers, as long as the AI tool or its output is used in the EU (Article 2(1) AI Act).

The problem is quite clear: precisely because they are general purpose, LGAIMs may be put to a thousand uses. At least one of them will practically always be high risk, as defined by the AI Act. For example, LGAIMs could be part of a decision engine used to rank and automatically reply to job candidates, asking them to supply further materials or to invite the top 100 candidates to the next round. Or they might draft patient letters in hospitals based on patient files, freeing up doctors' much-needed time for actual patient treatment. Such applications harbor great potential. At the same time, they also need regulatory oversight – but not the sort the AI Act envisages.

First, as noted, virtually all LGAIMs will qualify as high-risk systems under Article 4b AI Act due to their flexibility.[8)] As a consequence, a comprehensive risk management system needs to be established for all possible uses of the system – a task that borders on the impossible, given the amount of potential applications. The provider would have to list all possible applications, consider risks to all applicable fundamental rights and other legal positions of interest, and develop mitigation strategies for all of them. Ironically, the more powerful the model, the more unlikely it is that the provider will be able to comply with the AI Act, by virtue of the model's sheer versatility – there will just be too many scenarios to contemplate.

Second, some providers of LGAIMs are non-commercial open source developers. For them, at the very least, regulatory compliance with the generic risk management and other provisions of the AI Act will often prove prohibitively costly. Third, in conjunction with the recently published AI liability proposals, providers run a real risk of being held liable if their "high-risk" generative models make a "mistake", for example by committing an unrecognized translation error. Fourth, if the provisions deter providers from making LGAIMs available in the EU, downstream users seeking

to develop concrete applications will need to integrate less powerful, and presumably less rigorously vetted, AI tools into their decision-making systems – possibly making them less trustworthy and less safe. The rules may therefore undermine the very purpose of the AI Act. The AI Act section on GPAIS should, therefore, urgently be amended before enactment (see policy proposals below).

AI Content Moderation: The European Data Services Act

Things do not look much better when it comes to content moderation. The DSA, like the AI Act, applies to 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). It is the EU's new flagship legislation to mitigate harmful speech online, such as hate speech, fake news, or manipulative content. For example, platforms regulated by the DSA, such as Facebook or Twitter, need to establish a notice and action system under which users may flag potentially illegal content, which must then be reviewed and, if deemed illegal, deleted by the platform (Article 16 DSA). Larger platforms have to implement an entire internal complaint and redress system (Article 20 DSA) and offer out-of-court dispute settlement (Article 21 DSA), so that users do not have to go to court to challenge content. Accounts of repeat offenders must ultimately be suspended (Article 23 DSA). So-called trusted flaggers may register with Member States; content raised as problematic by them must be expeditiously reviewed (Article 22 DSA). Very large online platforms, in addition, need to establish a full-fledged compliance system, consisting, *inter alia*, of proactive risk management strategies and independent audits (Articles 33-37 DSA).

Is this of relevance for LGAIMs? It sure is. While the developers of ChatGPT have gone to great lengths to internalize content moderation by virtue of an AI supervisory instance blocking problematic requests and output (e.g., concerning discrimination or political violence), it is likely that this technical solution will never be watertight. Savvy users will find workarounds, using prompt engineering, as recent experiments show in which ChatGPT was convinced to launch a hate-filled shitstorm (the developers reacted promptly, though). Moreover, other LGAIMs do not feature these internal guardrails. The European Parliament is currently investigating Russian-based LENSA, which apparently enables rather unfettered, problematic image generation. Moreover, all LGAIMs, quite naturally, may equally be used to generate stories used for putting your kids to bed at night, as well as for launching mass-scale fake news attacks against democratic institutions, businesses, or individuals. The speed and syntactical accuracy of the generated content is rendered even more problematic, in this context, as users may prompt LGAIMs to adopt a seemingly knowledgeable, scholarly, or just highly confident tone. This makes them the perfect tool for the mass creation of highly polished, seemingly fact-loaded, yet deeply twisted fake news. Even for text generated by *bona fide* prompts, its factual accuracy may not match its astounding syntactical elegance. Hence, users must beware. And, of greater interest here, society must beware, particularly of users with malevolent intentions: 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.

Again, regulation is not ready to meet this challenge: the DSA only applies to so-called intermediary services (Article 2(1) and (2) DSA). These are conclusively defined in Article 3(g) DSA. They cover Internet access providers, caching services, and "hosting" services such as social media platforms. LGAIMs, arguably, do not qualify as either of these instances. Hosting services come closest, but they require the storage of information *provided by*, and at the request of, a user (Article 3(g) (iii) DSA). The trick with LGAIMs, however, is that the relevant content is decidedly not provided by the user, but by the LGAIM itself, having been prompted by the user via the insertion of certain query terms (e.g., "write an essay about content moderation in EU law in a lawyerly style"). With the DSA arguably inapplicable, the regulation of LGAIM content is left to the thicket of Member State speech regulation, which not only varies considerably across the EU, but also generally lacks the DSA instruments aiming to guarantee the speedy removal of harmful speech from the online sphere.

Of course, the DSA will apply if a user posts AI-generated content on a social network, such as Twitter. But at this point, it is often already too late to stem the tide of disinformation, hate speech, or manipulation. With LGAIMs, it is the creation that matters.[9)] Furthermore, the DSA generally does not regulate closed messenger groups on WhatsApp or Telegram (Recital 14 DSA), adding further to the deficiencies of reining in harmful AI content. If nothing is changed, the dark side of LGAIMs will go largely unaddressed.

Policy Proposals

How may sensible regulation look like going forward? Given the marked risks just sketched out, it seems unavoidable to regulate LGAIMs to a certain extent. But obligations must be tailored to their specific challenges, and compliance must be feasible for providers large and small alike. This is crucial not only for a healthy and competitive AI Act ecosystem, but also for environmental sustainability. Recent studies have shown that the impact of IT on climate change is real and significant, and AI's contribution is skyrocketing. Training, in particular of LGAIMs, is highly resource intensive. However, in the long run, these models may achieve significant sustainability advantages: they may be pre-trained once for, for example, 1000 different applications. In the alternative, 1000 less potent systems would have to be trained, quasi from scratch, generating potentially much higher greenhouse gas (GHG) emissions.

Against this background, we make three concrete, workable suggestions for LGAIM regulation: (i) transparency obligations; (ii) mandatory yet limited risk management; and (iii) expanded content moderation.

Transparency

First, LGAIMs should, irrespective of their qualification as high risk, be subject to two different transparency obligations. On the one hand, providers (e.g., LGAIM developers) should be compelled to report on performance metrics and any harmful content issues that arose during development (cf. Article 11, Annex IV AI Act). In addition, developers should also have to inform about the model's GHG footprint, to enable comparison and analysis by agencies, watchdogs, and other interested parties. Such disclosures could also be the basis for a Sustainability Impact Assessment for AI.[10)] On the other hand, users of AI-generated content should be under an obligation to disclose which parts of the content they make publicly available were produced by LGAIMs, or adapted from LGAIM-created content. As such a duty will be virtually impossible to enforce across the board, however, it would have to be backed up by technical strategies drawing on digital rights management. This may include conspicuous and inconspicuous watermarks woven by the LGAIM into its very content, in ways not easily removable by users. While particularly tech-savvy users may find a way to defeat them, average users may not. Currently, LGAIM developers are already exploring options to this effect, and the idea has been picked up by members of the European Parliament.

Risk Management

Second, risk management strategies, and other obligations reserved for high-risk applications in the AI Act, must be tailored to LGAIM realities. For instance, powerful models could be released in stages so that they can be thoroughly vetted. The full force of the high-risk section of the AI Act, however, should apply only if and when a certain LGAIM, or another GPAIS, is indeed used for high-risk purposes. For example, it would be the entity using the LGAIM for résumé screening that has to ensure compliance with the AI Act concerning this specific deployment – and not the often distinct LGAIM provider regarding the entire portfolio of possible purposes. This prevents prohibitive and inefficient regulation of the LGAIM at its source and tailors the obligations to the specific high-risk application for which it is used. Such a strategy corresponds to a well-known principle in product safety law. For example, screws in general do not have to be fit to hold together combat airplanes. Otherwise, nobody could afford simple screws for piecing together the latest IKEA unit. If, and only if, screws are used to manufacture combat airplanes, then indeed they have to comply with the strict regulations for products used in airborne military equipment. What is common sense in physical products should also apply to LGAIMs.

Content Moderation

Finally, third, the rules of the DSA must be adapted – *mutatis mutandis* – to generative AI systems. We need trusted flaggers, notice and action mechanisms, and the possibility of comprehensive audits for these models, just as we need them for social networks and other platforms. In fact, in the future, the dividing line between LGAIMs and platforms will probably become blurred, anyway, as Microsoft's pitch to invest another $10 billion in OpenAI, potentially for the integration of LGAIMs into its Bing search engine, clearly shows. Regulation needs to keep track, for the sake of the civility of our discourses, and to generate a level playing

field for developing and deploying the next generation of AI models, in the EU and beyond.

References

- In the following, we detail what prompts we used, here: What are large generative AI models?
- What distinguishes large generative AI models from other AI systems?
- Can you explain the technical foundations of large generative models in simple terms, so that an inexperienced reader understands it?
- What are the objectives, what are the obstacles when it comes to content moderation within large generative AI models?
- Ibid.
- How does content moderation work at ChatGPT?
- To reference the draft AI Act, we cite to the general approach adopted by the Council on December 6, 2022.
- The exception under Article 4c AI Act will generally fail due to the good faith restriction, Article 4c(2) AI Act.
- Note that we mean the creation of content in and for a specific use case.
- See also the recent push by MEPs for a Fundamental Rights Impact Assessment for AI, which would include an environmental dimension.
