The AI Act and the risks posed by generative AI models

Dag Elgesem*

[1]*Department of information science and media studies, University of Bergen*

Abstract

This position paper presents work in progress on the foundation of the not yet finalized AI Act and discusses whether the proposed AI Act has the resources to adequately identify and mitigate the risks posed by generative AI models like LLMs. All the details of AI Act are not yet decided but after the EU Council reached a consensus on a revised version of proposal issued by the Commission, the main principles of this risk-based model of regulation are in place. We argue that the methods for risk identification and mitigation in the act, while adequate to the management of many risks emerging from the use of AI systems, are not suitable for the management of all risks emerging from generative AI systems like LLMs

Keywords

AI Act, generative AI, language models, risk-based regulation, ethical risk evaluation

1. Introduction

In April 2021 the European Commission issued the draft proposal for the regulation of AI, the so-called AI Act [1]. In the public consultation on the draft proposal more than 300 stakeholders, including several of the tech industries and NGOs, issued position papers expressing different types of concerns. The European Council reached a compromise on a revised proposal for the AI Act in November 2022 and the European Parliament will vote on their position on the act in May this year. The process is expected to be completed and the AI Act turned into a law by the end of the year, after negotiations between the Council, the Parliament, and the Commission. Even though the process is not finished the main elements of the risk-based model for how to regulate AI are in place and it is therefore meaningful – and important – to critically discuss the overarching principles in the proposal. In this paper I will argue that the proposed regulation does not adequately handle the risks posed by generative AI like large language models (LLMs). As the basis of the argument, I will use the text of the most recent version of the regulation, issued by the European Council on November 25, 2022 [2].

NAIS 2023: The 2023 symposium of the Norwegian AI Society, June 14-15, 2023, Bergen, Norway

*Corresponding author.

dag.elgesem@uib.no (D. Elgesem)

0000-0002-0414-1885 (D. Elgesem)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3431/paper3.pdf

2. The risk model of the AI Act

The regulative model of the AI Act is risk-based and distinguishes between different levels of risk. Some types of AI applications are considered to pose unacceptable risks to individuals and are therefore prohibited under the proposed law. These include "AI-enabled manipulative techniques" [2], p. 25 that undermine people's autonomy by using subliminal techniques to persuade, deceive, or nudge people to act in different ways. Also, AI-systems are not allowed to be used for "social scoring", i.e., the use AI systems to collect and classify people's social behavior based on data collected across different social contexts. The reason is that such an application could lead to discriminatory and unfair treatment of individuals. The proposal then identifies several types of AI-applications that are characterized as "high risk". These include AI-applications used as part of products and might potentially harm the health and safety of consumers. But other types of applications are also considered to be "high risk".

> As regards high-risk AI systems other than those that are safety components of products, or which are themselves products, it is appropriate to classify them as high-risk if, in light of their intended purpose, they pose a risk of harm to the health and safety or the fundamental rights of persons, taking into account both the severity and the possible harm and its probability of occurrence, and they are used in a number of specifically pre-defined areas specified in the Regulation. [2], p. 35.

The regulation specifies these application areas in Annex III, and includes the use of AI in border control, education, recruitment, law enforcement, and critical infrastructure. The regulation also specifies criteria for the expansion of this list. In all the areas where the use of AI systems is classified as high-risks people are put in a vulnerable position in relation to the AI system with little control over the decisions that are made with the use of the systems. For such systems to be legal their development and application must comply with several requirements that are designed to minimize these risks.

> Requirements should apply to high-risk AI systems as regards the quality of data sets used, technical documentation and record-keeping, transparency and the provision of information to users, human oversight, and robustness, accuracy and cybersecurity. Those requirements are necessary to effectively mitigate the risks for health, safety and fundamental rights, as applicable in the light of the intended purpose of the system, and no other less trade restrictive measures are reasonably available, thus avoiding unjustified restrictions to trade. [2], p. 42

We see that AI-applications are classified as high-risk with reference to individual rights in clearly specified situations where these individuals are put in vulnerable positions with little control and that the providers and users of the applications are required to minimize the risk as they arise in these specific situations.

In addition to this, the regulation includes requirements of risk mitigation which apply to all AI systems, including a requirement of transparency of systems that involve interaction with "natural persons", whether they are classified as high-risk or not. "In particular, natural persons

should be notified that they are interacting with an AI system, unless it is obvious from the point of view of a natural person ...” [2], p. 55.

2.1. Ethical justification for the risk model

It is useful in the ethical evaluation of technological risks [3, 4] to consider the distribution of three different roles related to the use of technology. First, there is the question of who bears the cost associated with the risk, i.e., the risk exposed. Secondly, who benefits from the use of the technology and the risk? And, thirdly, who controls the technology and decides about the risks? These roles can be combined in different ways. The simplest and ethically usually least problematic are situations where one party occupies all roles: a person who uses a particular technology for her own benefit, shouldering all the consequences if something goes wrong. Conversely, the ethically most problematic situations are such that one party both benefits from the risk and decides whether to run it, while another party bears all the negative consequences. The situations identified as high-risk in the AI Act are of this latter kind, where individuals in vulnerable positions are subject to decisions with the use of AI-systems where other parties both decide about and benefit from the use of the technology and the involved risks. The measures the regulation require the users of the high-risk AI systems to put in place to mitigate the risks, are, essentially, ethical requirements to reduce the vulnerability of the risk exposed.

2.2. General purpose AI

General -purpose AI models that can be built into different types of application and used in different situations pose a challenge for this model. Should general models that could, potentially, be used in applications that are high-risk, themselves be classified as high-risk? In the public consultation following the Commission's original proposal several of the position papers issued by the tech industry, in particular Microsoft and Facebook, argued that would be very difficult for developers of these general models to take measures to mitigate the risks of possibly unknown downstream applications. Instead, they argued, the responsibility for mitigating risks should be put on those that deployed the high-risk applications using the general-purpose models in the situations specified in the regulation. The problem was addressed in the Council's revised version of the proposal, where a new section(12c) in the preamble on general-purpose AI was added. Here, the problem is recognized but not included in the text of the law.

> In particular, it is necessary to clarify that general purpose AI systems are AI systems that are intended by the provider to perform generally applicable functions, such as image/speech recognition, and in a plurality of contexts. They may be used as high-risk AI systems by themselves or be components of other high risk AI systems. Therefore, due to their particular nature and in order to ensure a fair sharing of responsibilities along the AI value chain, such systems should be subject to proportionate and more specific requirements and obligations under this Regulation while ensuring a high level of protection of fundamental rights, health and safety. [2], p. 23

While the Council chose to delegate the regulation to future implementation rules the EU Parliament, which is currently finalizing its treatment of the AI Act, has chosen to include the regulation of general-purpose AI in the act. According to reports, they will require developers of general-purpose AI models to provide the deployers of AI-systems that builds on the general models, with information necessary to enable the deployers to comply with the requirements for the mitigation of risks of high-risk systems.

3. The risk of generative models

We now turn to the central question of the paper: does the proposed regulation adequately manage the risks associated with generative AI models? The risks associated with these models have been discussed extensively in the months since the public release of ChatGPT and a call for a moratorium on the development of generative AI models have got the support of prominent figures in the research community [5]. The concerns have been of three types: 1) fear of losing control over a superior, general AI (long term risk), Some of the concerns are related to the fear of the long-term consequences, 2) fear of massive job losses of jobs (medium term risk), and 3) the spread of false information and toxic content (short-term risks). In the discussion below we will focus on the short-term risks, which are well documented. Studies have shown that LLMs can produce extremist content, giving fake yet convincing scientific and medical advice, and spread misinformation [6-11]. Some of the risks, e.g., using the technology to produce extremist or toxic content, arise from the potential for misuse by people with malicious intentions. But an important part of the risk stems from the LLMs' tendency for "hallucinating" convincing but completely made-up answers to queries [12,13]. The humanlike form of communication also contributes to the persuasiveness of the answers. The fear is that the extensive use of technologies that sometimes spread false but seemingly credible information will increase the spread of conspiracy theories, increase polarization, and erode trust in credible sources of information.

The challenge from the perspective of the AI Act is that, with its methods for risk identification the regulation would not recognize the generative AI applications as high-risk and the act's proposed measures for risk mitigation are inadequate for reducing the risks associated with these technologies. I will argue for the following claim:

The model for regulating the use of AI in the AI Act, based on minimizing risks in specified situation, is not adequate for regulating generative AI applications like LLMs, because 1) with such applications, all potential contexts of use and all relevant stakeholders cannot be specified, and the risks are substantial but hard to estimate, 2) the ethical justification for the principles for regulating "high risk application" in the AI Act does not extend to the evaluation of the risks that generative AI pose.

As pointed out above, the risk model of the AI Act identifies high-risk AI applications in relation to specific situations of use where individuals are put in a vulnerable position. Moreover, the risk in questions are risks of harm to individual rights. As an example, consider the justification for classifying systems used in the administration of education as high-risk:

> AI systems used in education or vocational training, notably for determining access, admission or assigning persons to educational and vocational training

> institutions or programmes at all levels or to evaluate learning outcomes of persons should be considered high-risk, since they may determine the educational and professional course of a person's life and therefore affect their ability to secure their livelihood. When improperly designed and used, such systems may violate the right to education and training as well as the right not to be discriminated against and perpetuate historical patterns of discrimination. [2], p. 36)

We see that the high-risk application is this area is identified in relation to specific situations where individuals are exposed to risks in virtue of their roles in the situations specified. Importantly for our discussion, the wider societal risks related to the technology are not considered with this method of risk identification. The risks associated that the generative AI models cannot not be captured in a similar way. Consider the problem of radicalization. A recent report, funded by OpenAI, documented how GPT3 could be used to generate highly convincing white supremacist and other extremist content [7]. The problem with a situation where someone uses this technology to produce extremist content is not primarily that that person's rights are at risk, but the risk that, through the spread of the material, other people can be radicalized. The risk situation is similar when the generative AI produces false information. That someone acquires false beliefs using generative AI is only part of the problem. The main problem is that, with the proliferation of false information in society, people could begin to distrust all information sources, not knowing if any sources or facts are credible. Over time, this could increase polarization and weaken the possibilities for rational discourse on important issues in society. Again, these are not risks to vulnerable individuals in specific roles but risks of harm emerging from the extended use of applications that spread disinformation across uses in different types of situations. It is thus not possible to identify the risks and who the risk exposed are, with reference to a particular type of situation. The method for risk identification in the AI Act will therefore not adequately identify and characterize these risks.

It might be suggested that the generative AI models are general purpose systems and the risk they pose can be handled though risk evaluations of the applications that are built on top of them. While it is true that many generative AI systems are general purpose systems, the risks of the systems that are built on them must, in the end, be identified and characterized with the methods discussed above. The risk assessment is transferred to the embedding applications, but the methods of risk identification will still be tied to specific contexts and the risks evaluated with reference to the way vulnerable individuals are treated with the use of AI application in those situations.

It is also clear that the requirement of transparency that applies to systems interacting with "natural persons" will not mitigate the risks posed by generative AI applications. As mentioned above, the requirement is that individuals that interact with AI systems shall be med aware that they are interacting with an artificial agent, unless it is obvious "from the point of view of a natural person who is reasonably well-informed, observant and circumspect taking into account the circumstances and context of use." [2], p. 55) It is of course possible that generative AI systems will be embedded in applications that deceive users into believing they are interacting with a human. But the risks we have seen so far emerging from the generative models do not stem from ignorance of the part of users of the fact that the answers are generated by an AI. Those that abuse the system to produce extremist or toxic content are of course fully aware

that they are using an AI system and disinformation can spread because the system provides convincing disinformation, not because people think they are interacting with a human.

Furthermore, one could ask if the generative AI models, in virtue of the risks they can give rise to qualify as posing unacceptable risks and therefore be prohibited under the AI Act? There are several reasons to think that they do not fall under the category of systems posing unacceptable risks. First, the treatment of the proposal for the AI Act in the EU parliament, which were not yet completed at the time of writing, was complicated by the impact of the release of ChatGPT. However, according to the 'Draft Compromise Amendments' [14] issued after the discussions in the two involved committees the Parliament will not propose to ban generative AI, which is characterized as a kind of foundational models:

> [F]oundational models should have information obligations and prepare all necessary technical documentation for downstream providers to be able to comply with their obligations under this Regulation. Generative foundational models should ensure transparency about the fact the content is generated by an AI system, not by humans. These specific requirements and obligations do not amount to considering foundational models as high risk AI systems, but should guarantee that the objectives of this Regulation to ensure a high level of protection of fundamental rights, health and safety, environment, democracy and rule of law is achieved. [14], p. 29

We see that generative models are classified in a lower risk category than high-risk systems, subject only to transparency requirements.

Second, those applications that are classified as posing unacceptable risks and therefore prohibited on the regulation, are either violating individual integrity (systems using subliminal and manipulative techniques), giving rise to discriminatory practices (social scoring), or surveillance. The risks posed by generative AI models, spreading of disinformation and toxic content, are different from any of these. Furthermore, generative AI models have a very large number of useful applications and have the potential to contribute significantly to innovations with social benefits, which it is a central goal of the AI Act is to facilitate.

My second argument for the claim that the risk management model in the AI Act does not adequately address the risks posed by generative AI models was that the ethical justification of the principles that underlie the AI Act does not extend to the evaluation of the risks that generative AI models pose. That ethical justification was based on a distinction between three different roles or positions in connection with the use of technologies involving risk: those who benefit from the risk, those who decide about the risk, and those that are exposed to the risk. The ethical argument for the model of identification and mitigation in the AI Act took as its starting point the observation that in the situations specified as high-risk in the AI Act, there was a challenging distribution of roles: the same actors that befit from using the risky AI systems is also deciding about the use of the systems, while a different actor is exposed to the risks that the systems pose and will suffer harm if they fail. This is an ethically challenging type of situation because of the vulnerability of those that are risk exposed, the power relation between the decision maker and the risk exposed, and the incentive the decision maker could have to exploit the risk exposed. In some cases, the risk exposed could also to some extent benefit from

the use of the technology, if the use of the AI system makes an application process in which the risk exposed is involved more efficient, for example. But unless the distribution of roles is unchanged the ethically challenging nature of the situation does not change significantly.

The regulation requires that the party that benefits from the risk and decides about the risk take the cost of putting in place measures to reduce the risk of the risk exposed to an acceptable level. Given the asymmetry of the situation and that the risk exposed often will not have the freedom to choose not to be exposed to the risk, this is a fair distribution of costs and responsibilities. With the generative AI models, however, the same ethical problem remains, with one party both benefiting from the risk and deciding whether to run it, while a different party, the risk exposed, bears the risk. But with the generative models the relationship of the risk exposed to the technology is different from those targeted by the AI Act. The crucial difference is that the risks do not arise primarily in the situation of use but in the context of communication in the wider social contexts of which the user is a member. As argued above, with applications like LMMs, where the risks are related to such things as the spread of false information, radicalization, and erosion of trust, the risk situations cannot be specified in the same way as with the AI Act because these risks arise across a broad range of different situations of use and potentially affect a broad and unspecified range of stakeholders. The measures the act prescribes for the mitigation of the risk will therefore not resolve the ethical challenge in this case.

4. Conclusion

We have argued for the negative conclusion that, with the resources provided by the risk model in the AI Act, the ethical challenges of general-purpose cannot be adequately met. This does not show, of course, that regulations that mitigate the risks associated with these models cannot be found. But such regulations will have to force the developers of the systems to make sure that they are mostly truthful and prevent the systems from spreading disinformation. Whether this level of control is possible to achieve with LLMs seems to be an open question [15].

5. References

1 European Commission. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. 2021.

2 Council of the European Union. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. 2022.

3 S.O. Hansson, Ethical Risk Analysis, in: The Ethics of Technology, edited by S.O. Hansson, Rowman and Littlefield, London, 2017.

4 J.Wolff, Five Types of Risky Situations, Law, Innovation and Technology, 2 (2), 2010.

5 J. Bengio, Joshua et al. Pause Giant AI Experiments: An Open Letter. Future of Life Institute, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

6 L. Weidinger et al., Taxonomy of Risks posed by Language Models. FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3531146.3533088

7 T. Brown et al. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs] (July 2020). http://arxiv.org/abs/2005.14165 arXiv:2005.14165.

8 D. Nozza, F. Bianchi, and D. Hovy. 2021. HONEST: Measuring Hurtful Sentence Completion in Language Models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, 2398–2406. https://doi.org/10.18653/v1/2021.naacl-main.

9 J.W.Rae et al. Scaling Language Models: Methods, Analysis and Insights from Training Gopher. arXiv:2112.11446 [cs], 2021.

10 S. Lin, J. Hilton, and O. Evans. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv:2109.07958 [cs], 2021. http://arxiv.org/abs/2109.07958 arXiv:2109.07958

11 R. Zellers, A. Holtzman, Hannah R., Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. 2020. Defending Against Neural Fake News. arXiv:1905.12616 [cs], 2020. http://arxiv.org/abs/1905.12616 arXiv:1905.12616

12 K.McGuffie and A. Newhouse, The Radicalization Risk Posed by GPT-3 and Advanced Neural Language Models. Middlebury Institute of International Studies at Monterey, 2019.

13 C. Schyns, The Lobbying Ghost in the Machine, Corporate Europe Observatory, 2023

14 European Parliament, Proposal for a regulation of the European Parliament and of the Council on harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, Draft Compromise Amendments on the Draft Report, 2023.

15 S.R.Bowman, Eight Things to Know about Large Language Models,arXiv:2304.00612 [cs.CL], 2023