ChatGPT and the AI Act

Helberger, N.; Diakopoulos, N.

DOI
10.14763/2023.1.1682
Publication date
2023
Document Version
Final published version
Published in
Internet Policy Review

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 12 **Issue** 1

ESSAY

OPEN ACCESS

ChatGPT and the AI Act

Natali Helberger *University of Amsterdam*
Nicholas Diakopoulos *Northwestern University*

DOI: https://doi.org/10.14763/2023.1.1682

Published: 16 February 2023
Received: 15 February 2023

Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Helberger, N. & Diakopoulos, N. (2023). ChatGPT and the AI Act. *Internet Policy Review*, *12*(1). https://doi.org/10.14763/2023.1.1682

Keywords: ChatGPT, AI governance, Artificial intelligence, Digital Services Act (DSA), Generative AI

Abstract: It is not easy being a tech regulator these days. The European institutions are working hard towards finalising the AI Act in autumn, and then generative AI systems like ChatGPT come along! In this essay, we comment the European AI Act by arguing that its current risk-based approach is too limited for facing ChatGPT & co.

It is not easy being a tech regulator these days. The European institutions are working hard towards finalising the AI Act in autumn, and then generative AI systems like ChatGPT came along. A powerful language model, trained on unprecedented amounts of data and able to engage in astonishingly diverse conversations – from writing movie reviews and poems to grading school essays, judging resumes or writing software code. Across a range of use cases and contexts, you ask in natural language and you get a smooth-sounding answer. Millions of people are already using it. And it's for exactly these reasons that generative AI like ChatGPT challenges the current risk-based approach in the AI Act.

Released only barely a couple of months ago, the imagination of what the technology can do and mean for society and its values are running high, also among legislators. Last week, the European Parliament reportedly suggested, as a last-minute addition, to expand the potential reach of the AI Act by placing AI-generated texts that could be mistaken for human-generated and deep fakes that say or do things that never happened to the list of high-risk categories. At the same time, new language in the regulation itself includes a definition of general-purpose AI (which includes systems such as ChatGPT and Dall-E) and that the European Commission will lay down further details on how the high-risk provisions apply in an implementing act. The question is: does the AI Act's risk-based approach still fit the case of generative AI?

We argue that generative AI systems such as ChatGPT differ on at least two important points from the 'traditional' AI systems the Act has originally been written for: dynamic context and scale of use. Generative AI systems are not built for a specific *context* or conditions of use, and their openness and ease of control allow for unprecedented *scale* of use. The output of generative AI systems can be interpreted as media (text, audio, video) by people with ordinary communication skills, lowering, therefore, significantly the threshold of who can be a user. And they can be used for such a variety of reasons to some extent because of the sheer scale of extraction of data that went into their training. Three hundred billion words for ChatGPT alone, spanning all kinds of contents available on the internet – from personal data to policy documents, news reporting, literary texts and art.

These characteristics – no intended purpose and scale of adoption – challenge the current approach in the AI Act in at least three important ways: the feasibility of sorting generative AI systems into high/no high risk categories, the unpredictability of future risks, and concerns around private risk ordering.

High-risk, no-risk, general-risk

According to the current logic of the AI Act, the categorisation of an AI system as high or no risk depends on the purpose of use that the provider envisages. All systems that are intended to be used in one of the areas specified in annex III of the regulation are considered high risk. In all other situations, AI systems fall under the no-risk category (or, in the case of deep fakes and chatbots: a low-risk category with transparency as the ultimate regulatory answer). But for general purpose AI, it is not the provider but rather the *professional user* who determines how they will use the system. It is the user who determines whether the system falls into the low or high-risk category. Some of the risks for society will result from the way *end* users, aka "consumers" use these systems. Depending on how the phrase 'may be used as high risk AI systems' in the new Article 4b of the AI Act is interpreted, this could mean that the legal obligations for high-risk AI only take effect once the generative AI is being used in a high-risk area.

From the point of view of society and fundamental rights, this is too late. The whole point about generative AI as a general-purpose AI system is that because they can be used for so many different purposes, it is paramount to incentivise the *providers* of systems to think about the safety of these systems from the onset, starting with the difficult question of data quality. Otherwise, any potential biases, privacy violations, unlawful uses of content or other instances of unfairness in the data or the model will trickle down into a myriad of possible future applications. Under the current version of the AI Act, the incentives for providers to do so are potentially close to zero. The AI Act entitles the provider to 'opt out' from applying the high-risk provisions by explicitly excluding all high-risk uses in the instructions to use (Art. 4 (c) AI Act). And it is yet entirely unclear what the obligations of professional or end users are in such a situation. End users in particular are largely excluded from the scope of the AI Act.

The alternative scenario would be that all generative AI systems would fall under the high risk category because it cannot be excluded that they *may* be used also in a high-risk area. In that case, there may be a serious risk of over-regulation.

For this reason, rather than trying to fit general-purpose AI systems into existing high-risk categories, we propose that they should be **considered a general-risk category in their own right,** similar to the way that chatbots and deep fakes are considered a separate risk category of their own, and subject to legal obligations and requirements that fit their characteristics. For example, right now, the data management obligations in the AI Act address mainly concerns around completeness,

accuracy, and lack of bias. With generative AI and the pure scale of extraction of training data from all kinds of sources, lawful or not and authorized or not, far broader questions of what we call *extraction fairness* come to the fore. Extraction fairness is about broader legal and moral concerns regarding the large-scale exploitation of training data without the knowledge, authorisation, acknowledgement or compensation of their creators.

From pre-defined risks to systemic risk monitoring

With 100 million active users in the first months after its launch, ChatGPT has been described as the "fastest-growing consumer application ever launched". The pure scale of adoption, in combination with the versatility and general purpose characteristics of the technology, challenge the AI Act's risk-based approach in a second important way: it is simply impossible to predict if, and if so, what the risks are that we can expect from unleashing extremely powerful AI models on society. The EU parliament's recent additions to the AI Act frame the central risk of generative AI systems as of lack of authenticity. But authenticity is not necessarily the main challenge for health, safety, and the realization of fundamental human rights. How about risks to privacy or creativity, a risk that arises not from the inauthenticity of the information but from factual mistakes, overreliance on e.g. the legal expertise of ChatGPT, the lack of verifiability and ease to scale up the amplification of disinformation, issues of cybersecurity, or a melt-down of regulatory authorities because the sheer scale of operations confronts them with insurmountable enforcement challenges. We have yet to find out. Society is only beginning to explore what widely accessible generative AI systems are capable of, in hackathons, in school classes, at work, or in the living room.

So instead of hastily trying to solidify the risk areas in a difficult-to-change Annex III, we need to think in more dynamic ways of monitoring and mitigating any risks for individuals and society. The systemic risk monitoring approach in art. 34 of the Digital Services Act (DSA) could be an inspiration. Under the DSA, Very Large Online Platforms and Very Large Search Engines are already obligated to monitor their algorithmic systems regularly for any actual and foreseeable negative effects on fundamental rights and societal processes, including such that arise from the implementation of generative AI models. It is conceivable that a comparable obligation to monitor for and mitigate systemic risks on a regular basis should also apply to the providers of very large generative AI models.

Private ordering

Another consequence of the general purpose character of generative AI models is that the intended purpose and conditions of use are ultimately defined in the (contractual) relationship between user and provider. According to the logic of the AI Act, a central element in the provider-user relationship are the instructions from the provider to the user. In a situation where the actual use and intended purpose are not generally foreseeable, instructions will play an even more important role in outlining the safety requirements and conditions of lawful use for each use case. The instructions, and more generally, the contractual relationships between provider and user, will also be critical in properly allocating responsibilities and cooperation obligations. Users (professional and end users) may depend for compliance with their legal obligations (e.g. about data quality or human oversight) quite critically on cooperation from the provider, and vice versa. Providers rely on users to share their experiences with the system to further improve and use the systems responsibly to help make them safe.

This also means that the quality and fairness of the contractual arrangements between users and providers of generative AI will play a decisive role in addressing, allocating and mitigating (systemic) risks for society. For example, Open AI, the provider of ChatGPT, is currently stipulating its own 'greater risk' categories, not all of which are covered by the high-risk categories of the AI Act (including, e.g. the use for healthcare, therapy, wellness, coaching, finance, news). For these categories, the company imposes on users a distinct set of obligations. The provider of another generative AI system, Stable Diffusion, formulates an entire battery of other prohibited uses of the system that go far beyond art. 5 AI Act, including bans on generating false information to harm others, defamation or harassment, or providing medical advice and medical results interpretation.

These contractual attempts at concretising responsibility and rules of engagement are useful and important in contributing to emerging social practices of responsible use of generative AI. Having said so, as with any contractual terms and usage restrictions, they are subject to typical concerns about private ordering, such as information asymmetries, unequal negotiation powers and legal incentives to limit contractual liability at the costs of the weaker party (see already the newly proposed art. 4 c of the AI Act, above). A complex system of private ordering could also defy the broader purpose of the AI Act to promote legal certainty, foreseeability and standardisation. In its current form, the fairness and quality of the user-provider instructions are not subject to any requirements, and the end user is even almost entirely absent. Our third recommendation would be to include mecha-

nisms of regulatory scrutiny regarding the fairness, quality and adequacy of contractual terms and instructions.

In conclusion

In this essay, we have reflected on the suitability of the AI Act in its current form to adequately deal with general-purpose AI and, more specifically, generative AI systems such as ChatGPT. We conclude that generative AI challenges some of the core concepts and principles of the AI Act in its current form. In particular, the current risk-based approach needs to be better suited to the lack of a pre-defined purpose and risks related to the scale of use and extraction of training data.

Updating the AI Act to include generative AI adequately is more than just fitting generative AI into the current high-risk provisions. Instead, we argue in favor of considering generative AI and general purpose AI more broadly as a general-risk category in its own right. We also point to the need for critical scrutiny and recalibration of the Act's obligations in light of the characteristics of generative AI systems, and how those obligations apply to the complex interplay between the different actors (providers, professional users and end users). This includes a proposal to consider a general monitoring obligation for systemic risks, similar to the approach taken under the DSA, as well as more attention to the (contractual) relationship between providers of generative AI systems and their users, professional and end users. More generally, there is a need to think through the distinction between provider, professional user, and end user. It is laudable that the Act in its current version foresees more time and discussion to decide how exactly to include generative AI, and general purpose AI systems more generally under its scope. However, we should also be aware that the regulatory challenges that generative AI raises go far beyond mere questions of implementation, and the choices we make in that process will have far-reaching societal and economic implications.

Published by
Alexander von Humboldt Institute for Internet and Society

in cooperation with
CREATe
R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya
cnrs centre internet et societe
University of Tartu Johan Skytte Institute of Political Studies