Why the European AI Act transparency obligation is insufficient

With the development of the AI Act, the EU is making the first and globally most ambitious attempt to regulate AI. However, the proposed AI Act, which employs a risk-based taxonomy for AI regulation, encounters difficulties when applied to general-purpose Large Language Models (LLMs) and likely underestimates the risks posed by these new AI models. This correspondence warns that the AI Act must evolve further to mitigate such risks.

A main challenge is that LLMs like ChatGPT generate unverified information and produce fictitious content with confidence. For example, ChatGPT can generate pertinent, but non-existent academic reading lists [1]. Data scientists explain that such effects are caused by "hallucination" [2] and because LLMs function like "stochastic parrots" [3]. Hallucination occurs when LLMs generate text based on their internal logic or patterns, rather than the true context, leading to confidently but unjustified and unverified deceptive responses. LLMs are called stochastic parrots as they repeat training data or its patterns, rather than actual understanding or reasoning.

LLMs produce text by reusing, reshaping, and recombining the training data in new ways to answer new questions while ignoring the problem of authenticity and trustworthiness of the answers. Although most answers are of high quality and true, the content of the answers is fictional. Even though most training data is reliable and trustworthy, the recombination of this data into new answers in a new context may lead to untrustworthiness, as the trustworthiness of information is conditional and often depends on context. If this precondition of trustworthy data disappears, trust in answers will be misplaced. While the LLMs' answers may seem highly relevant to the prompts, they are made-up.

Merely improving the accuracy of the models through new data and algorithms is insufficient, because the more accurate the model is, the more users will rely on it, and thus be tempted not to verify the answers, leading to greater risk when stochastic parrots and hallucinations appear. The risk is beyond measure if users encounter these problems in especially sensitive areas such as healthcare or the legal field. Even if utilizing real-time internet sources, the trustworthiness of LLMs may remain compromised, as exemplified by factual errors in new Bing's launch demo [4].

These risks can lead to ethical concerns, including misinformation and disinformation, which may adversely affect individuals through misunderstandings, erroneous decisions, loss of trust, and even physical harm (e.g., in healthcare). Misinformation and disinformation can reinforce bias, [5] as LLMs may perpetuate stereotypes present in their training data.

In the proposed taxonomy of the European AI Act, LLMs could on the one hand be categorized as high-risk AI due to its generality, but this may impede EU's AI development. On the other hand, if general-purpose LLMs are regarded as chatbots they fall within the limited-risk group. But merely imposing transparency obligations (i.e., providers need to disclose that the answer is generated by AI) would be insufficient [6]. Users should be clearly informed when they are interacting with AI, but they also need to be able to assess the reliability and trustworthiness of LLMs' answers, to distinguish between truth and made-up answers. When a superficially eloquent and knowledgeable chatbot generates unverified content with apparent confidence, users may trust the fictitious content without undertaking verification. Therefore, the AIA's transparency obligation is not sufficient.

Additionally, the AIA does not address the role, rights, or responsibilities of end-users. As a result, they have no opportunity to contest or complain about LLMs. Moreover, the AIA does not impose any obligations on users while the occurrence and spread of disinformation is largely due to deliberate misuse by users. Without imposing responsibilities on the user side, it is difficult to regulate the harmful use of AI by users.

Apart from the AIA, the Digital Service Act (DSA) aims to govern disinformation. However, the DSA's legislators only focus on the responsibilities of the intermediary, overlooking the source of the disinformation. Imposing obligations only on intermediaries when LLMs are embedded in services is insufficient, as such regulation cannot reach the underlying developers of LLMs. Similarly, the Digital Markets Act (DMA) focuses on the regulation of gatekeepers, aiming to establish a fair and competitive market. Although scholars recently claim that the DMA has significant implications for AI regulation [7], the DMA primarily targets the effects of AI on market structure; it can only provide limited help on LLMs. The problem that the DSA and DMA will face is that both only govern the platform, not the usage, performance, and output of AI *per se*. This regulatory approach is a consequence of the current platform-as-a-service (PaaS) business model. However, once the business model shifts to AI model-as-a-service (MaaS) [8], this regulatory framework is likely to become nugatory, as the platform does not fully control the processing logic and output of the algorithmic model.

Therefore, it is necessary to urgently reconsider the regulation of general-purpose LLMs [9]. The parroting and hallucination issues show that minimal transparency obligations are insufficient, since LLMs often lull users into misplaced trust. When using LLMs, users should be acutely aware that the answers are made-up, may be unreliable, and require verification. LLMs should be obliged to remind and guide users on content verification. Particularly when prompted with sensitive topics, such as medical or legal inquiries, LLMs should refuse to answer, instead directing users to authoritative sources with traceable context. The suitable scope for such filter and notice obligations warrants further discussion from legal, ethical and technical standpoints.

Furthermore, legislators should reassess the risk-based AI taxonomy in the AIA. The above discussion suggests that the effective regulation of LLMs needs to ensure their trustworthiness, taking into account the reliability, explainability and traceability of generated information, rather than solely focusing on transparency. Meanwhile, end-users, developers and deployers' roles should all be considered in AI regulations, while shifting focus from PaaS to AI MaaS.

Zihao Li[1,2]
[1]CREATe Centre, School of Law, University of Glasgow
[2]Stanford Law School, Stanford University

Competing interest statement
The author declares no competing interests.

Acknowledgement
The author would like to express sincere gratitude to Professor Martin Kretschmer for his valuable insights, comments, and support during the development of this paper. Our discussions have played a pivotal role in the formation of this paper. This work was supported by Modern Law Review Scholarship.

Reference

1. https://teche.mq.edu.au/2023/02/why-does-chatgpt-generate-fake-references/
2. Ji, Z. *et al.* Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* (2022). doi:10.1145/3571730
3. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the Dangers of Stochastic Parrots. in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (ACM, 2021). doi:10.1145/3442188.3445922
4. Kan, M. Demo of Microsoft's AI-Powered Bing Included Several Small Mistakes. *PC Magazine* https://www.pcmag.com/news/demo-of-microsofts-ai-powered-bing-included-several-small-mistakes (2023).
5. van Bekkum, M. & Zuiderveen Borgesius, F. *Comput. Law Secur. Rev.* **48**, 105770 (2023).
6. Edwards, L. *The EU AI Act: a summary of its significance and scope*. (2022).
7. Hacker, P., Cordes, J. & Rochon, J. Regulating Gatekeeper AI and Data: Transparency, Access, and Fairness under the DMA, the GDPR, and Beyond. *SSRN Work. Pap.* (2022). doi:10.2139/ssrn.4316944
8. Sun, T., Shao, Y., Qian, H., Huang, X. & Qiu, X. Black-Box Tuning for Language-Model-as-a-Service. in *Proceedings of the 39th International Conference on Machine Learning* 20841–20855 (PMLR, 2022).
9. Hacker, P., Engel, A. & List, T. Understanding and Regulating ChatGPT, and Other Large Generative AI Models. *VerfBlog* (2023). doi:10.17176/20230120-220055-0.