European Journal of Risk Regulation (2024), **15**, 719–736
doi:10.1017/err.2024.4

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

"More than Words": A Legal Approach to the Risks of Commercial Chatbots Powered by Generative Artificial Intelligence

Sara Migliorini

Faculty of Law, University of Macau, Macau, China
Email: sara.migliorini@eui.eu

Abstract

The recent commercial release of a new generation of chatbot systems, particularly those leveraging Transformer-based large language models (LLMs) such as ChatGPT, has caught the world by surprise and sparked debate about their potential consequences for society. While concerns about the existential threat posed by these technologies are often discussed, it is crucial to shift our focus towards the more immediate risks associated with their deployment. Such risks are further compounded by the lack of proactive measures addressing users' literacy and the for-profit model via which these chatbots are distributed. Drawing on research in computer science and other fields, this paper looks at the immediate risks triggered by these products and reflects on the role of law within a broader policy directed at steering generative artificial intelligence technology towards the common good. It also reviews the relevant amendments proposed by the European Parliament to the European Commission's proposal for an AI Act.

Keywords: Artificial intelligence; EU AI Act; foundation models; generative AI; regulation of emerging technology

I. Introduction

In September 2022, OpenAI, a charity that later added a for-profit entity to its governance structure,[1] released a free version of a chatbot system named ChatGPT, which it then turned into a pay-for subscription plan. Since then, ChatGPT's successors, and other similar products,[2] have gotten the world talking about their shocking capabilities. This new wave of commercial chatbots also prompted a debate on the possibility that humanity may be getting closer to a new, more powerful type of artificial intelligence (AI)[3] and on all its potentially disruptive effects on our society – from the job market to education and beyond.

[1] N van der Horst, "Embedding checks and balances in steward ownership: the case of OpenAI" (*Transformative Private Law Blog*, 11 December 2023) <https://transformativeprivatelaw.com/embedding-checks-and-balances-in-steward-ownership-the-case-of-openai/>.

[2] For a review of all similar available products and their capabilities, see, for example, EM Humphries et al, "What's the Best Chatbot for Me? Researchers Put LLMs through Their Paces" (*Nature*, 27 September 2023) <https://www.nature.com/articles/d41586-023-03023-4>.

[3] S Bubeck et al, "Sparks of artificial general intelligence: early experiments with GPT-4" (2023) arXiv preprint arXiv:2303.12712.

This paper focuses on this new generation of chatbots released commercially at the end of 2022 and during 2023 and more generally discusses all current and future chatbots exploiting Transformer-based large language models (commonly abbreviated as "LLMs"). With the commercial release of these products, humans started to wonder whether this impressive technology questions our place in the world and whether a future in which humans would be obsolete is approaching at a faster pace than we could have ever anticipated.[4]

Yet, the fear that these machines will bring about the end of human civilisation as we know it, and other dystopian and eerie scenarios,[5] obfuscate the more imminent risks that are associated with the underlying technology.[6] Some such risks may have already occurred and become more severe because these chatbots have been made available to a wide share of the population without any prior actions being taken to addres the literacy of users and via a for-profit model. And while this technology is still in the "hope and hype" phase, now is the appropriate time for lawyers and policymakers to take a hard look at it and act to steer it towards the common good and away from risks that can already be foreseen or imagined.

This paper explains how the new generation of chatbots works (Section II) and what actual risks for humans, society and the planet appear to be associated with them (Section III). It then looks at how the legal system should respond to such potential risks and discusses possible regulatory choices, with a special focus on the proposal for a European Union (EU) regulation on AI,[7] currently under discussion (Section IV). Section V concludes.

II. How chatbots using large language models work

To reflect on the risks stemming from new technologies, and their potentialities as well, we first need to understand their inner workings. Indeed, AI-based systems such as LLMs are often perceived as "black boxes": the users only see that they provide an input to the system, which in turn produces an output. The chatbots considered in this article work exactly in this way: the user can provide instructions or a question and they will receive a reply in written form from the chatbot. However, ignoring how the input is processed and how the output is composed contributes to some of the risks associated with these systems, namely the risks associated with their distribution amongst a vulnerable and naïve population of users, who have not been given any information regarding how the system works. This section offers a brief presentation of the technology powering most of the available commercial products, such as ChatGPT and Bing Chat, drawing from writings in the fields of computer science.

[4] Under the influence of the industry, part of the public debate about generative AI concentrated on the "existential risk" linked to this technology. See K Roose, "A.I. Poses 'Risk of Extinction', Industry Leaders Warn" (*The New York Times*, 30 May 2023) <https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html>.

[5] These fears are not new but have accompanied AI since its beginning. See IJ Good, "Speculations Concerning the First Ultraintelligent Machine" (1965) 6 Advances in Computers 31, 33, as reported by A Plasek, "On the Cruelty of Really Writing a History of Machine Learning" (2016) 38(4) IEEE Annals of the History of Computing 6.

[6] EM Bender et al, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (2021) Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 610.

[7] European Commission, "Proposal for a Regulation Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act)" COM (2021) 206 final (21 April 2021; hereinafter, the "EU AI Act"); and "Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts", COM(2021)0206 – C9-0146/20212021/0106(COD) (hereinafter the "EU AI Act – Parliament Amendments").

At the outset, it needs to be noted that LLMs are classified as a type of "foundation model". This term was introduced in 2021 to describe AI-based systems that are trained on broad data and can be adapted to a wide range of downstream tasks.[8] The term "foundation" also means that, although these pieces of software are unfinished, insofar as they need adapting to a specific task, they do provide the basis for many different applications. The term "foundation" therefore conveys the importance of the correct development and deployment of such software. The term has become influential in the literature since its introduction. The European Parliament (EP) has indeed proposed a few amendments to the AI Act to take this term into account and the category of software that it defines. In particular, Article 3(1)(c) AI Act – Parliament Amendments defines a "foundation model" as "an AI system model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks". Although foundation models can be applied to tasks such as vision and robotics, this paper focuses specifically on those powering chatbots and therefore those allowing the computer-based processing of natural language (commonly abbreviated "NLP").

NLP is an area of computer science as old as the research on "artificial intelligence" itself, if not older.[9] The idea of using machines for translating texts in a different language, and generally processing natural language, followed the successful attempts at code cracking during World War II.[10] Nonetheless, it was not until large numbers of texts became available in a digital format that the performance and potentialities of NLP were significantly augmented.[11] Through advancements in the field of machine learning and neural networks from the 2010s until today, modern LLMs have thrived and evolved beyond anything that existed beforehand.

In essential terms, chatbots based on LLMs generate texts in response to instructions typed into the chatbot (called "prompts") by using statistical techniques. The way in which LLMs are trained allows them to compose texts based on the most statistically probable association of words that follow each other in human-generated texts. In order to achieve this, NLP tools of the modern era are powered by three essential features.[12] The first one is the possibility of being trained on immense amounts of text available online and in a digital, machine-readable format, covering the span of human knowledge and the many human ways of retelling our experiences of the world. Secondly, a new way of doing machine learning followed from the introduction of an infrastructure called the Transformer by Google,[13] which is a new type of architecture of neural networks, and its descendants, such as BERT.[14] Finally, and crucially, the past few years have also brought advanced computational capabilities thanks to extremely powerful hardware that is able to process a multitude of complex calculations simultaneously.

The way in which an LLM is able to "understand" and "learn" from a pre-existing text in a digital format, and the way in which it "knows" words in a given language, is extremely

[8] R Bommasani et al, "On the Opportunities and Risks of Foundation Models" (2021) arXiv preprint arXiv:2108.07258.

[9] CD Manning, "Human Language Understanding & Reasoning" (2022) American Academy of Arts & Sciences <https://www.amacad.org/publication/human-language-understanding-reasoning>; Bender et al, supra, note 6.

[10] Manning, supra, note 9, 127.

[11] ibid, 129.

[12] D Luitse and W Denkena, "The Great Transformer: Examining the Role of Large Language Models in the Political Economy of AI" (2021) 8 Big Data & Society <https://doi.org/10.1177/2053951721104773>.

[13] Introduced for the first time in A Vaswani et al, "Attention Is All You Need" (No arXiv:1706.03762, arXiv, 1 August 2023) <http://arxiv.org/abs/1706.03762>.

[14] See J Alammar, "The Illustrated Transformer" <https://jalammar.github.io/illustrated-transformer/>. BERT stands for "Bidirectional Encoder Representations from Transformers", an encoder-only architecture. See J Devlin, MW Chang, K Lee and K Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (No. arXiv:1810.04805v2, 11 October 2018) <https://arxiv.org/abs/1810.04805v2>.

complex for a layperson to fully understand, but anyone should be able to grasp the essentials of how these tools work, particularly the lawyers and policymakers who are called to set and apply the rules by which these tools are deployed in society.

In simplified terms, LLMs are composed of a series of twenty to thirty building blocks, called encoder or decoder blocks, depending on the model, each of which is made of two different layers. The first layer is based on a feature called "self-attention", which performs two tasks. Firstly, it transforms a word into an integer number (called *a token*), which, from that moment on, will represent that word in the LLM. Secondly, it assigns each word with "some" meaning by linking the word itself to the context in which it is found in the training sentences. This operation in practice transforms the number representing the word into a matrix of numbers that all represent the word at all the possible positions and combinations with other words that the LLMs have found in the training data (an operation called "*embedding*"). The magnitude of embedding is only limited by the capabilities of the hardware and the data used in the training. It therefore becomes clear why the availability of large amounts of text in a digital format and advanced hardware capabilities have powered the rise of LLMs.

The other layer of each building block of the LLM is able to produce an output based on the input that it receives from the preceding block, which is a possible "answer" to the instruction. This output passes to the following block. Each block is able to improve and refine the LLM's "understanding" of the word in this same way. At the moment of projecting the final output, the LLMs use another neural network to proceed to transform the matrix of numbers into a probabilistic calculation of which of the tokens (ie words) that it contains should be selected as a reply to the question asked in the chatbot. The token selected according to probabilistic calculations as the most appropriate is re-transformed in the corresponding word and projected by the LLM in its natural language reply. This operation is done for every word or combination of words that constitutes the natural language text answering the question in the chatbot.

A modern-day LLM first goes through a pre-training phase, during which it is exposed to an enormous amount of text, spanning different topics and styles. The purpose of this phase is to create an LLM that has a large vocabulary (and the corresponding *embedding*) that is not related to a particular field or a particular purpose. During pre-training, the model, by definition, will not be able to predict the statistically most likely word to follow a certain sentence. Errors in prediction are spotted and then the correct result is fed into the LLM, so that it may learn. This operation is done millions of times during the pre-training phase. In addition, these LLMs are capable of performing self-supervised learning, which dispenses with human oversight during pre-training and seems to have proven to be even more effective than human-supervised training.[15]

The technical process through which the data have been selected for the training of commercially available chatbots, such as ChatGPT, is generally known to researchers and experts in the field. For example, ChatGPT's foundational model has been trained on a freely accessible library of texts, called the "Common Crawl".[16] The data in the Common Crawl have been recorded from webpages since 2008 via tools called "web crawlers" that are able to covertly scrape information from websites without leaving a trace of their presence.[17] A group of researchers has applied three different filters to the Common Crawl to produce a database called "C4.EN.NOCLEAN", which contains more than 2.3 TB of text, to

[15] A Radford et al, "Language models are unsupervised multitask learners" (2019) 1 OpenAI Blog 8.

[16] T Brown et al, "Language models are few-shot learners" (2020) 33 Advances in Neural Information Processing Systems 1877, at 1884.

[17] This database is called the "Common Crawl" <https://commoncrawl.org/the-data/get-started/>.

exclude, for example, some offensive words.[18] In the training of ChatGPT, this "clean"[19] version of the Common Crawl has been complemented and filtered with the addition of a few other datasets.[20] On the one hand, two Internet-based datasets containing books, called "Books1" and "Books2", whose content is unspecified by OpenAI but has been documented in previous publications, seem to contain a vast array of academic publications from, for example, PubMed, along with material from YouTube and a "mix of fiction and non-fiction books".[21] On the other hand, OpenAI added to ChatGPT's training some Internet-based sources made out of scraped materials, such as English-language Wikipedia pages and another freely available dataset known as "Webtext".[22] For the earlier version of ChatGPT released at the end of 2022, we know that all of these datasets have been used but also that the training gave more prominence to the "clean" Common Crawl.[23] Although this explanation may be enough to grasp the technicality of the selection of data, it provides the layperson, the consumer and society at large with little meaningful information regarding the actual content of the information used to train these LLMs and the effects that such content has on the output of the commercially available chatbot systems.[24]

After the pre-training phase, the LLM goes through another phase, called *fine-tuning*, during which the model can be trained for a specific task – for instance, academic writing or translation. During this phase, pre-trained models undergo a new type of training that uses labelled data taken from a more specific dataset that is adapted to the specific task to which the fine-tuning is aimed. In this phase, the datasets include both the input and the wanted corresponding outputs. This allows the pre-trained model to "learn" to generate outputs that are increasingly similar to the labelled data provided. This operation can also be done via a special machine learning technique called "*reinforcement learning*", whereby the model is rewarded when it produces an output that is sufficiently similar to the desired one. During this process, humans can be involved in different ways, including via real-time interaction with the model. The exact data used for fine-tuning are generally not known for most of the commercially available chatbots.[25]

III. Potential risks associated with the development and commercial deployment of chatbots using large language models

This section of the paper attempts to explain the possible negative impacts on society of the new generation of chatbots. At this stage, these risks appear possible, and therefore worthy of consideration, to avoid sleepwalking into a future in which humans are made worse off by the introduction of these new technologies. Some of the risks discussed in this section have already been realised to various degrees. Conversely, some other risks may turn out to be less daunting than they appear now. The objective of this section is to present the risks that seem plausible whilst taking into consideration the current

[18] C Raffel et al, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" (2020) 21 Journal of Machine Learning Research 5485.

[19] For a critique of the biases in datasets, see A Caliskan et al, "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases" (2017) 356(6334) Science 183; CC Perez, *Invisible Women: Data Bias in a World Designed for Men* (New York, Abrams 2019).

[20] Brown et al, supra, note 16, 1884.

[21] L Gao et al. "The pile: An 800gb dataset of diverse text for language modeling" (2020) arXiv preprint <https://doi.org/10.48550/arXiv.2101.00027>.

[22] Compiled by Aaron Gokaslan and Vanya Cohen in 2019 and available open source at <https://skylion007.github.io/OpenWebTextCorpus/>.

[23] Brown et al, supra, note 16, figure 2.2 at 9.

[24] This problem has been flagged by experts and will be discussed in Section III.

[25] For example, at this stage, we do not have information about the data used in the fine-tuning of ChatGPT-4.

deployment at scale and for commercial purposes in an untrained and unprepared population. Section IV will then reflect upon the areas of law that need to address such risks in order to mitigate them or prevent them altogether.

For clarity, the potential risks are divided into (1) risks related to the input provided to chatbots, (2) risks related to the output of chatbots and, finally, (3) systemic risks associated with the deployment of these tools for commercial purposes.

1. Risks related to the training of chatbots

The first set of risks has to do with the way in which the LLMs that underpin the newly available chatbots are trained. As explained in Section II, the training of LLMs comprises a pre-training phase and a fine-tuning phase. Both phases rely on the possibility to train the language model on very large amounts of data, the content of which is not transparent.[26] This way of training raises a series of questions.

Firstly, the use of large datasets that have been created starting from libraries of crawled data and then refined using different filters, some of which are unaccounted for, raises the question of whether information fed to LLMs during training is biased, to the disadvantage of different groups in society. In every system that relies on machine learning techniques, "datasets form the critical information infrastructure underpinning [machine learning] research and development, as well as a critical base upon which algorithmic decision-making operates".[27] Notwithstanding the crucial role played by datasets in any machine learning application, such as LLMs, work that relates to datasets is heavily under-incentivised as opposed to work focusing on the development of more efficient algorithms.[28] Institutions of research in the field of computer science, comprising industry and academic institutions, seem to feed into this lack of recognition of the valuable work that would be necessary to curate databases,[29] worsening the status of such crucial infrastructure for machine learning. As a consequence, publications accompanying new datasets have been found to under-specify the decisions that go into the collection, curation and annotation of datasets,[30] leading to a lack of transparency and reliance on best practices regarding the curation of datasets and no general interest in whether the datasets are reliable in the first place.[31] In turn, this vicious circle feeds a phenomenon that has been called the "naturalisation of datasets": as the datasets used for LLMs become increasingly well-known and relied upon on a routine basis by industry and researchers, the history behind their creation is lost, "in a manner that ultimately renders the constitutive elements of their formation invisible".[32] This lack of documentation of the process and content behind the datasets used in the training of LLMs is alarming per se, and it prompts the question regarding accountability for the content produced by

[26] See supra, Section II, notes 13ss and accompanying text.

[27] E Denton et al, "On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet" (2021) 8 Big Data & Society <https://doi.org/10.1177/20539517211035955>.

[28] RS Geiger et al, "Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?" in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, Association for Computing Machinery 2020) p 325 <https://dl.acm.org/doi/10.1145/3351095.3372862>; see also ES Jo and T Gebru, "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning" in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, Association for Computing Machinery 2020) p 306 <https://dl.acm.org/doi/10.1145/3351095.3372829>; B Hutchinson et al, "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, Association for Computing Machinery 2021) p 560 <https://dl.acm.org/doi/10.1145/3442188.3445918>.

[29] Hutchinson et al, supra, note 28.

[30] Geiger et al, supra, note 28.

[31] ibid.

[32] ibid.

commercial products, such as ChatGPT3. Ample literature has documented the existence of LLMs' biases and discriminatory outputs. For example, LLMs have been found to associate the word "Muslim" with violence.[33]

These risks linked to discrimination and exclusions in algorithms and machine learning systems using large and opaque datasets have been widely known for years,[34] and very little has been done since then to mitigate them and develop different ways to harness the potentialities of AI. In the era of the commercialisation of chatbots exploiting LLMs, it is high time for policymakers and the legal system to find the right tools to prevent discriminatory outputs, starting with streamlining the process of training and the choice of data.

Secondly, another concern that arises is whether the information fed to LLMs via large datasets used for training can be used at all for such a purpose. At least two problems can be flagged under this perspective. On the one hand, large datasets harbour a real risk of exposure of the personal data of unknowing individuals. It has been proven that, under some conditions, it is possible to reverse engineer data present in large datasets used for training in order to extract personal data referring to identifiable individuals.[35] This is a way by which the personal data of individuals that are available on the Internet can be retrieved by third parties. As it has been explained, the fact that such data were already publicly available on the Internet does not in itself warrant authorisation or give consent to further processing.[36] This way of retrieving the personal data of individuals can also expose identified people to harm, or more generally to unwanted attention.[37] On the other hand, some of the information used to train models, such as English-language texts used to train LLMs, may have been put on the Internet with the assumption that they would not be used for such a purpose, or they may have been put on the Internet at a time when this particular type of use was not known and thus surely not contemplated by their authors. Although these circumstances do not in themselves demonstrate that the use of this type of text is a violation of existing laws on copyright or authorship or contractual arrangements linked to websites, they do raise the question of how to control for such possibilities if the datasets used are opaque and non-transparent.

Finally, another very alarming risk concerns the exploitation of workers in the Global South, who are called to work on the process of fine-tuning by labelling data and other tasks related to reinforcement learning. For example, reporting has uncovered such practices being used by OpenAI in Kenya.[38]

In addition to these risks, the legal system and policymakers need also to reckon with the fact that the training of LLMs happens mainly in a non-supervised way and thus without human oversight. This in turn renders the question of how to think about accountability for any possible violation of laws or harm that is caused during or by the training process. More generally, the way in which LLMs work does not leave much room for inquiry and thus reinforces the "black box" model.

[33] A Abid et al, "Large Language Models Associate Muslims with Violence" (2021) 3(6) Nature Machine Intelligence 461.

[34] For example, regarding machine learning algorithms running on social media, see SU Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York, NYU Press 2018), and more generally the works referred to by Bender et al, supra, note 6, at 613–15.

[35] N Carlini et al, "Extracting Training Data from Large Language Models" (2020) arXiv e-prints arXiv-2012; H Li et al, "Multi-Step Jailbreaking Privacy Attacks on ChatGPT" (2023) arXiv preprint arXiv:2304.05197.

[36] See, for example, the recent joint statement on data scraping and the protection of privacy by twelve national data protection authorities (24 August 2023) <https://www.priv.gc.ca/en/opc-news/speeches/2023/js-dc_20230824/#fn1>.

[37] A Alanwar et al, "Data-Driven Reachability Analysis From Noisy Data" (2023) 68 IEEE Transactions on Automatic Control 3054.

[38] B Perrigo, "OpenAI Used Kenyan Workers on Less than $2 Per Hour to Make ChatGPT Less Toxic" (*Time*, 18 January 2023) <https://time.com/6247678/openai-chatgpt-kenya-workers/>.

2. *Risks associated with the outputs of chatbots*

The second set of risks that can be identified relates to the content produced by chatbots leveraging LLMs. As was explained previously, the newly released commercial chatbots generate content in the form of text as a response to the instructions input by the user ("prompt").

In this respect, the first broad question that arises is whether the output of the chatbot may harm humans in any way. A non-specific answer to this question would be that there seem to be instances in which harm is not only a plausible risk but is already established. As mentioned in Section III.1, biases in the training data have translated into discriminatory outputs. As an example of the harm directly caused by chatbots, the Italian Data Protection Authority (DPA) has ordered an urgent temporary limitation on the processing of personal data relating to users located in Italy by the company operating Replika, an AI-powered chatbot generating a "virtual friend".[39] The Italian DPA has found, via some tests and other evidence regarding replies generated by Replika, that the chatbot posed risks to minors and, generally speaking, "emotionally vulnerable individuals". With a similar decision, the Italian DPA has also blocked ChatGPT for a few weeks pending explanations and commitments from OpenAI regarding the processing of personal data of Italian users, especially minors.[40] It is also worth noting that the draft EU AI Act, currently under discussion, prohibits the commercialisation of AI-powered tools that can manipulate users or otherwise exploit the vulnerabilities of minors and other groups.[41] These few examples seem sufficient to establish that the output produced by chatbots can harm humans, especially minors or other groups of vulnerable individuals.

Nonetheless, when considering the potential for harm arising out of the new generation of chatbots, attention should also be paid to studies that have highlighted everyone's risk of exhibiting some vulnerability that can be exploited, including by AI. It has been put forward by literature in the field of behavioural economics and anthropology that everyone, immersed as we are in an "endless chain of acts of consumption", becomes a vulnerable consumer.[42] The overwhelming nature of the demands that the consumer market puts on humans fosters a mindset of scarcity, whereby mental space for certain cognitive tasks is absorbed by other issues, putting consumers in a situation that is structurally vulnerable vis-à-vis their counterparts.[43] This is particularly true for individuals in situations of poverty or marginalisation, but it remains a valuable point for the vast majority of consumers. In such a situation, humans become "disengaged" consumers and "find themselves in vulnerable purchasing situations, not because of particular cognitive failings or socio-demographic characteristics, but because the structure of the consumer markets on which they evolve leads to apathy through obfuscation".[44] Based on these ideas, the new generation of chatbots, deployed at scale for commercial purposes, may have the potential to harm everyone, to the extent that they find themselves, at different moments throughout their lifetime and even throughout their day, in a situation of scarcity, disengagement and, thus, vulnerability. In addition, as

[39] The decision of the Italian DPA is available in English at <https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9852214>.

[40] The decision of the Italian DPA on ChatGPT is available at <https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870832>.

[41] RJ Neuwirth, "Prohibited Artificial Intelligence Practices in the Proposed EU Artificial Intelligence Act (AIA)" (2023) 48 Computer Law & Security Review 105798.

[42] D Lunn, "Are Consumer Decision-Making Phenomena a Fourth Market Failure?" (2015) 38 Journal of Consumer Policy 315.

[43] S Mullainathan and E Shafir, *Scarcity: Why Having Too Little Means So Much* (1st edition, New York, Times Books, Henry Holt and Company 2013).

[44] P Siciliani et al, *Consumer Theories of Harm: An Economic Approach to Consumer Law Enforcement and Policy Making* (Oxford, Hart Publishing 2019).

demonstrated by countless experiments, not only children or other vulnerable groups are potential victims of unethical or illegal marketing techniques that exploit subliminal or similar techniques.[45] It cannot be assumed that the risks associated with the output of chatbots, especially with respect to their effects on the human mind, can be purely avoided by protecting special categories, such as minors. A deeper and interdisciplinary reflection is needed in order to understand the exact scope of the potential harms for individuals and society at large.

The second set of risks is associated with the accuracy of the outputs. The first experiences with ChatGPT3 and Bing Chat have reported mixed results. In some instances, the replies of the chatbots are well written and factually accurate, such as when asked to summarise a given paragraph. On the other hand, other studies have highlighted the presence of gross inaccuracies and plain falsehoods within the replies of ChatGPT.[46] If allowed to circulate (eg in the form of social media posts), these falsehoods and inaccuracies may raise issues in any sphere of social life, from politics to health and safety. In addition, as mentioned previously, commercial products such as ChatGPT3 have been allowed to be deployed at scale and for commercial purposes in an untrained population, which in vast part has not been prepared to deal with this technology. In addition, most of the commercial chatbots that have been released so far do not seem to provide sources for their statements. Accordingly, users have no means of verifying whether the information given is reliable or not. And if a user is required to double-check every piece of information that emerges from the chatbots, their utility may be greatly undermined.

A third broad set of questions, which is linked to the problems of training and the lack of sources as highlighted above, is whether the output of chatbots might interfere with the rights of human authors and creators. In particular, outputs can interfere with copyright and other rights attached to human creativity and also constitute plagiarism in society at large and within the narrower field of education. As it has been argued, "ChatGPT's ability to produce large amounts of plausible-sounding content and to rewrite existing text in different styles, making plagiarism detection near-impossible, may stretch the current system to its limits and undermine trust".[47] Conversely, a claim has been made that some LLM-based tools may be able to detect whether a text has been written by another LLM model, although doubts remain regarding their efficacy.[48] Although this is crucial, detecting such practices would only be the first step in finding a solution to the problem of preserving human creations from the outputs of chatbots without obviously jeopardising the great support that tools such as ChatGPT could provide to authors and creators in general.

And, in this respect, it is also possible that text-generating AI-based tools will augment the creative potential of humans in the same way as other AI-based tools have done in other instances. For example, after an AI-based system had defeated the human world champion in the game of Go, a board game similar to chess,[49] human professional Go players started training and playing games against AI-based systems. Ultimately, a player who had been training with AI beat an AI-based system.[50] Research in cognitive psychology

[45] RJ Neuwirth, *The EU Artificial Intelligence Act. Regulating Subliminal AI Systems* (London, Routledge 2022).

[46] See, for example, S Hargreaves, "'Words Are Flowing Out Like Endless Rain Into a Paper Cup': ChatGPT & Law School Assessments" (2023) (2023-03) *The Chinese University of Hong Kong Faculty of Law Research Paper*.

[47] Editorial, "The AI Writing on the Wall" (2023) 5(1) Nature Machine Intelligence 1.

[48] R Williams, "AI-text detection tools are really easy to fool" (*MIT Technology Review*, 2023) <https://www.technologyreview.com/2023/07/07/1075982/ai-text-detection-tools-are-really-easy-to-fool/>.

[49] J Somers, "How the Artificial Intelligence Program AlphaZero Mastered Its Games" (*The New Yorker*, 28 December 2018) <https://www.newyorker.com/science/elements/how-the-artificial-intelligence-program-alphazero-mastered-its-games>.

[50] K Southern and L Angeles, "Man Beats Machine at Go Thanks to AI Opponent's Fatal Flaw" (*The Times*, 14 December 2023) <https://www.thetimes.co.uk/article/man-beats-machine-at-go-thanks-to-ai-opponents-fatal-flaw-nc9vqmrvf>.

has submitted that training with AI-based computers fostered human Go players' ability to think outside the box and made them better decision-makers, allowing them to eventually outsmart the machine.[51] Therefore, it is possible that generative AI tools, such as chatbots, could be used to foster human abilities and skills, "augmenting" our potentialities in a fruitful collaboration between humans and machines. It will then be crucial to correctly recognise and protect the rights of "augmented human creators" as well as to clarify the role and rights of the programmers and owners of the AI tools used and possibly of the AI creator as well.[52]

3. Systemic risks associated with the deployment of the new chatbots at scale for commercial purposes

As was mentioned previously, NLP and machine learning are not new techniques. In particular, chatbots have been deployed for years – for example, in customer support functions. Yet, as mentioned above, the current wave of new chatbots is different. Firstly, the capabilities of the new LLM-powered chatbots are greatly increased thanks to the advances in computing and the vast availability of datasets for training. Although these advances are impressive and should be welcomed, a lot remains to be done in terms of the energy consumption of the required investments. Secondly, these chatbots are now sold as commercial products, and they have reached a vast and untrained population. As pointed out above, the majority of users possess limited information on how chatbots work and are not aware that the logical statements made by chatbots do not follow the rationality commanding human language and meaning and instead follow a logic based merely on statistical reasoning. In addition, the general public knows little or nothing about the training of ChatGPT and similar products and what parameters and constraints guide them, although some such information may be accessible to experts in the field. Furthermore, when a user gets a reply from these chatbots, they are usually not provided with the sources of the information it contains. Thus, confirming the accuracy of a statement provided may be burdensome for the average user, which, given the lack of AI literacy of most users, will probably lead to the general acceptance of the chatbot's statement as true.

All of these factors contribute to the potential risks of harm to individuals and society at large. As a consequence, regulators should be able to identify and prevent a series of risks that are linked to these factors. Building on writings from computer science and other disciplines, this paper identifies three categories of such types of risks.

Firstly, policymakers and lawyers should urgently address the environmental costs of training and operating these chatbots. The impressive escalation in the amount of computing used to train and operate LLMs has a significant environmental impact. As pointed out previously, machine learning is an energy-hungry endeavour, which translates notably into CO_2 emissions, one of the main drivers of climate change.[53] Additional research has also studied the impacts of machine learning in general on all greenhouse gas (GHG) emissions, identifying three different stages of machine learning in which high

[51] M Shin et al, "Superhuman artificial intelligence can improve human decision-making by increasing novelty" (2023) 120(12) Proceedings of the National Academy of Sciences of the United States of America e2214840120.

[52] N Lucchi, "ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems" (2023) European Journal of Risk Regulation 1 (footnote 85) doi: 10.1017/err.2023.59. See also S Sachs, "US Copyright Office Rules AI-Generated Artwork, Content Not Legally Protected" (*The Hill*, 24 February 2023) <https://thehill.com/homenews/3872614-us-copyright-office-rules-ai-generated-artwork-content-not-legally-protected/>. For the first decision and story, see B Edwards, "Artist Receives First Known US Copyright Registration for Latent Diffusion AI Art" (*Ars Technica*, 22 September 2022) <https://arstechnica-com.cdn.ampproject.org/c/s/arstechnica.com/information-technology/2022/09/artist-receives-first-known-us-copyright-registration-for-generative-ai-art/amp/>.

[53] Bender et al, supra, note 6, 612–13.

levels of GHG emissions are involved: computing-related impacts, the immediate impacts of applying machine learning and system-level impacts.[54]

Additionally, research has shown that institutions and stakeholders in the field of machine learning tend to concentrate on the optimisation of models rather than operating a whole cost–benefit analysis of a new, more powerful technology with respect in particular to its environmental costs and energy efficiency.[55] Although we should welcome calls to use AI and machine learning in the context of mitigation and adaptation efforts regarding climate change,[56] a more holistic approach to the costs and benefits of this technology appears to be the essential first step to be performed.[57] Considerations related to climate change need to inform all policy decisions regarding LLM deployment, and it has to become a pivotal objective, at the policy and legal level, to rein in energy-costly models.

An additional and very pressing negative effect of the environmental costs of machine learning and chatbots in particular is that such costs tend to accrue to disadvantaged groups in society, which are not the same groups that benefit from the financial or social advantages of the technology and are in general subject to many different instances of discrimination and environmental racism.[58] With the deployment of a new generation of chatbots at scale and the profits generated by their operating companies, the issue of representation of marginalised groups within the decision-making processes leading to ever-bigger models with higher energy consumption levels and emissions should be high on the agenda of policymakers at the national and international level.

Secondly, it is foreseeable that the risks that the new generation of chatbots pose to individuals – identified in the previous sections as discrimination of certain groups, loss of privacy, interference with creative rights, misleading statements and other manipulation risks – will be amplified at the societal level by the sheer number of users of such chatbots. In short, when a chatbot is used by a million users every single day, harm to individuals may become harmful to society. Let us imagine that it becomes possible to extract the personal data of individuals from the replies of one of these chatbots.[59] If this happens to one person, it is a data breach and a privacy intrusion relative to such an individual. If this happens to millions of individuals, the problem becomes a cybersecurity issue and needs to be addressed at the societal level. Similar reasoning can be applied for all of the above risks.

Finally, a pressing systemic issue is the disruption effect that these chatbots may provoke in many of the fundamental social institutions that underpin liberal democracies: the job market, the education system, the political system and the maintenance of free competition. The increasing availability of commercial products running on LLMs, which can generate output that is overall as good as human output, may prompt companies to reduce their number of employees.[60] Similarly, education institutions across all grades may find it difficult to continue to teach and assess students within the traditional

[54] LH Kaack et al, "Aligning Artificial Intelligence with Climate Change Mitigation" (2022) 12 Nature Climate Change 518.

[55] R Schwartz et al, "Green AI" (2020) 63 Communications of the ACM 54.

[56] D Rolnick et al, "Tackling Climate Change with Machine Learning" (2023) 55 ACM Computing Surveys 1.

[57] See, for example, the research of M Treviso et al, "Efficient Methods for Natural Language Processing: A Survey" (2022) <https://arxiv.org/abs/2209.00099>.

[58] Bender et al, supra, note 6, 612–13.

[59] M Nasr et al, "Scalable extraction of training data from (production) language models" (2023) arXiv:2311.17035. See also press reports of a similar problem that was spotted regarding OpenAI: L Bernardone, "ChatGPT Suffers First Major Data Leak: Systems Taken Down after Bug Exposes Payment Info" (*Information Age, ACS*, 28 March 2023) <https://ia.acs.org.au/article/2023/chatgpt-suffers-first-major-data-leak.html>.

[60] The World Economic Forum estimated that 85 million jobs will be replaced by AI in the next decade, and, although more jobs will be created, certain categories and groups will not benefit from them, unless policymakers intervene. See World Economic Forum, "The Future of Jobs Report 2020" (2020) <https://www.weforum.org/reports/the-future-of-jobs-report-2020/>.

curriculum when students have access to these tools.[61] Chatbots may also produce social media posts or other types of scripts that can convey false or misleading information and be diffused at scale amongst the population, with the potential to disrupt democratic processes and free elections.[62]

All of these risks of disruption must be taken into account by policymakers and correctly addressed using old and new legal tools to allow society to benefit from – rather than be overwhelmed by – chatbots running on LLMs.

IV. Large language models and the law

As explained in Section III, the new generation of chatbots and generative AI in general have and will continue to have significant repercussions across many different sectors of society and, consequently, many different subfields of law. At the time of writing, policymakers and regulators at the national and supranational level are debating and putting forward ideas regarding whether and how they should regulate generative AI.[63] The remainder of this paper provides some reflections regarding how policymakers can think about the law as a means within this endeavour and reflects upon the current solutions adopted by the EU AI Act – Parliament Amendments.

1. Old and new questions for the law

A first issue that has arisen in the policy debate at the national level about the best way to regulate generative AI is whether new regulation is necessary at all and if new regulation could hamper innovation, putting the national economy at a disadvantage as compared to other countries that may let the new technology run free of regulation. However, from a legal point of view, it appears that this should not be the first question that policy should address. On the contrary, there should first be a reflection on what existing laws that are enforceable at present are relevant and applicable to generative AI.[64]

Along this line, in many jurisdictions there is a significant body of enforceable legal rules that should be relevant and applicable to many aspects of the deployment of chatbots. Under this perspective, these amazingly disruptive tools do not raise disruptive legal questions but rather old ones. For example, as discussed earlier, the Italian DPA has applied the General Data Protection Regulation (GDPR) to the chatbots Replika and

[61] See, for example, a study by UNESCO on the several challenges posed by generative AI to the education system: E Sabzalieva and A Valentini, "ChatGPT and Artificial Intelligence in Higher Education: Quick Start Guide" (UNESCO, 2023) <https://etico.iiep.unesco.org/en/chatgpt-and-artificial-intelligence-higher-education-quick-start-guide>.

[62] See, for example, recent reports: J Haidt Schmidt, "AI Is About to Make Social Media (Much) More Toxic" (*The Atlantic*, 5 May 2023) <https://www.theatlantic.com/technology/archive/2023/05/generative-ai-social-media-integration-dangers-disinformation-addiction/673940/>; C Klein, "'This Will Be Dangerous in Elections': Political Media's Next Big Challenge Is Navigating AI Deepfakes" (*Vanity Fair*, 6 March 2023) <https://www.vanityfair.com/news/2023/03/ai-2024-deepfake>.

[63] See, for example, A Bradford, "The Race to Regulate Artificial Intelligence" (*Foreign Affairs*, 23 June 2023) <https://www.foreignaffairs.com/united-states/race-regulate-artificial-intelligence>.

[64] As reported widely, the US Federal Trade Commission has made it clear that AI-based tools do not benefit from an exemption from the relevant rules. See, for example, C Lima, "Regulators pledge to use 'laws on the books' to tackle AI abuses" (*Washington Post*, 26 April 2023) <https://www.washingtonpost.com/politics/2023/04/26/regulators-pledge-use-laws-books-tackle-ai-abuses/>. See also Consumer Financial Protection Bureau, Department of Justice's Civil Rights Division, Equal Employment Opportunity Commission and Federal Trade Commission of the U.S., Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems (25 April 2023) <https://files.consumerfinance.gov/f/documents/cfpb_joint-statement-enforcement-against-discrimination-bias-automated-systems_2023-04.pdf>.

ChatGPT and found them in breach of those existing rules.[65] Another relevant area of law is competition law. As it has been maintained, the introduction of the Transformer, the hardware underpinning LLMs, has led to a concentration of the required immense computing in the hands of a few companies or states around the world.[66] A very topical question is how to ensure that such a concentration of the means of computing does not lead to a fragmentation of the market and the creation of an oligopoly that prevents smaller players from accessing it. Existing rules of competition should apply to similar situations and behaviours of enterprises. For example, China's policymakers are launching initiatives to allocate computing power to different services and areas to exploit efficiently the country's existing resources and allow different market players to access state-owned or privately owned computers.[67] In a neighbouring field, rules of consumer protection should also be relevant to products like ChatGPT.[68] Finally, rules that seek to protect individuals and groups from discrimination should be fully applicable and applied whenever generative AI is used.[69]

Undoubtedly, existing laws that are relevant to generative AI will need to be adapted and tweaked, at least partially, to meet some of the specific challenges raised by this new technology. In other instances, it will be necessary to clarify the extent to which existing laws do apply to generative AI. These adaptations and tweaks may happen either when existing rules are applied by courts or authorities in specific cases or in a preventative way by lawmakers amending existing laws.

A parallel could be drawn in this respect with how tax and labour laws have been applied to platforms allowing peer-to-peer economic exchanges, such as Airbnb and Uber. In matters of taxation, it was unclear whether existing laws would apply to peer-to-peer, short-term rentals made through Airbnb, especially concerning tourist taxes that, in major cities around the world, local administrations impose on tourists and that are usually collected by hotels and operators of other traditional forms of touristic accommodation rentals. Such taxes, along with many other aspects of the economic activity allowed by Airbnb, have been regulated in major tourist cities after such cities experienced the negative consequences of the rise in such types of rentals for tourists,[70] in particular with respect to the payment by hosts of tourism taxes.[71] To achieve this shift in tax rules,

[65] S McCallum, "ChatGPT Banned in Italy over Privacy Concerns" (*BBC News*, 31 March 2023) <https://www.bbc.com/news/technology-65139406>.

[66] Luitse and Denkena, supra, note 12.

[67] This project goes under the name of the National Unified Computing Power Network (NUCPN) and not only has the means to address the competition issue but also and foremost has been created to face the export controls on critical components that have targeted Chinese companies – for example, see the report at R Creemers et al, "Is China's Tech 'Crackdown' or 'Rectification' Over?" (*DigiChina*, 25 January 2023) <https://digichina.stanford.edu/work/is-chinas-tech-crackdown-or-rectification-over/>. China also wants to build a "China Computing NET(C^2NET)" to establish a national strategy that "channels computing resources from the east to the west" (东数西算战略) and to systematically optimise the computing infrastructure layout as mentioned in the 2023 Digital China Plan. See "加快打造全国算力"一张网"_滚动新闻_中国政府网" 人民日报 (27 March 2022) <https://www.gov.cn/xinwen/2022-03/27/content_5681738.htm>; 观察者, "全国一体化算力算网调度平台正式发布，已接入天翼云、华为云、阿里云等" (6 June 2023) <https://www.guancha.cn/politics/2023_06_06_695599.shtml>.

[68] This has recently been stressed by the US Federal Trade Commission; see supra, note 64. China also rolled out administrative measures on generative AI in August 2023 that provide for the protection of privacy and safeguard the rights of consumers and users: <http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm>.

[69] This has recently been stressed by the US Civil Rights Division of the Department of Justice (see supra, note 64), and China's regulation on algorithmic recommendation stipulates that algorithmic recommendation service providers shall take measures to prevent the dissemination of harmful online content. See Provisions on the Administration of Algorithm-generated Recommendations for Internet Information Services <https://www.gov.cn/zhengce/2022-11/26/content_5728941.htm>.

[70] D Von Briel and S Dolnicar, "The Evolution of Airbnb Regulation – An International Longitudinal Investigation 2008–2020" (2021) 87 Annals of Tourism Research.

[71] ibid, 3.

different regulatory techniques have been deployed: some cities have collaborated with the platforms and drafted common guidelines (eg Barcelona),[72] and others have introduced legally binding rules (eg Tokyo).[73]

Similarly, Uber did not consider drivers using its platform as employees but rather as self-employed. Consequently, Uber would not respect the obligations of an employer, such as paid holidays and proper breaktimes. It has been through litigation that, under certain conditions, Uber drivers have been recognised to be employees.[74]

More generally, what we are currently witnessing in the field of chatbots and generative AI is a transitional phase, during which these clarifications and adaptations are happening gradually, as society and the regulators realise the challenges linked to the deployment of the technology, similar to other fields that have been characterised by innovation.[75] During this phase, it is important for policymakers to clearly state that chatbots do not benefit from exemptions and that existing laws apply insofar as relevant, including data protection laws and fundamental rights. In this respect, the recent Chinese rules for the regulation of generative AI are interesting, as they clarify that any such product needs to respect all existing laws, in addition to the few specific rules introduced by such measures.[76] In addition, those who act as legal advisors of operators and users of chatbots and other generative AI tools should be mindful of the possible legal risks, in particular with respect to the possible application of laws already in force to the uses of this new technology.

After having surveyed the existing laws that constrain AI, there will remain other important and truly innovative questions that will not be well apprehended by existing laws. In these respects, the issue will indeed be whether new rules are needed and what form such rules should take: top-down regulation, judge-made law or various forms of soft laws and collaborative rules. One field that seems ripe for profound modifications is copyright, both regarding the use of copyrighted works in the training of the models and regarding the protection of work generated by authors and artists with the support of generative AI tools. These issues seem truly novel, in the sense that the current legal framework seems unable to correctly apprehend them. It therefore seems that new ways of rewarding authors for allowing the training of generative AI on their works, along with the opening of the possibility to protect creative work that uses generative AI, should be considered as possible developments of the legal system.

2. *The draft European Union Artificial Intelligence Act*

At the time of writing, the EU's lawmakers are discussing an ambitious, comprehensive regulation on AI.[77] The Commission's proposal was published in the spring of 2021, with the desire to position the EU as the world regulator of AI, "winning" the global regulatory race.[78] However, given the many issues raised by AI, especially with respect to

[72] ibid, 3.

[73] ibid, 3.

[74] See cases in the Netherlands and England: MA Russon, "Uber Drivers Are Workers Not Self-Employed, Supreme Court Rules" (*BBC News*, 19 February 2021) <https://www.bbc.com/news/business-56123668>.

[75] E Biber et al, "Regulating Business Innovation as Policy Disruption: From the Model T to Airbnb" (2017) 70 Vanderbilt Law Review 1561, 1567; S Fredman and D Du Toit, "One Small Step towards Decent Work: Uber v Aslam in the Court of Appeal" (2019) 48 Industrial Law Journal 260, 262.

[76] See supra, note 64.

[77] For the full references to the Commission's proposal and European Parliament's amendments, see supra, note 68.

[78] A Bradford, *Digital Empires: The Global Battle to Regulate Technology* (Oxford, Oxford University Press 2023) pp 324–59; NA Smuha, "From a 'Race to AI' to a 'Race to AI Regulation': Regulatory Competition for Artificial Intelligence" (2021) 13(1) Law, Innovation and Technology 1, 21–25.

fundamental rights, intensive discussions have taken place in the EP during the legislative procedure. The rise of chatbots at the end of 2022 further complicated the legislative process and prompted the EP to add amendments specifically targeting foundation models. The EP's amendments are discussed in this section, albeit the latest discussions and available information seem to confirm that most of these amendments will not be included in the final text of the AI Act.[79]

At the outset, it is interesting to note that the EP has embraced the transitional nature of legal rules regarding foundation models adopted at this early stage for the commercial deployment of this technology. Recital 60(h) of the EU AI Act – Parliament Amendments states that, since foundation models are a "new and fast-evolving" AI application, the Commission and other specialised EU bodies should "periodically assess the legislative and governance framework of such models". Secondly, the added recitals also show that, in proposing specific rules for foundation models, the EP was moved by two main concerns. On the one hand, foundation models are instrumental to many different products ("downstream applications and systems"),[80] and therefore their correct deployment is necessary to avoid a negative "domino effect" regarding such products. On the other hand, the EP has expressed the willingness to protect providers of AI products that rely on a foundation model, which they did not develop, trying to ensure that such providers receive from the developers of the foundation model all of the necessary information and support to ensure compliance of their downstream applications with the future AI Act.[81] Finally, it appears that many of these concerns, which are expressed in detail in the recitals, did not necessarily find a specific corresponding rule in the amendments to the text of the regulation itself. For example, the reference to foundation models provided through API[82] is only found in the recitals.[83]

The main provision proposed by the EP on foundation models is Article 28b, complemented by the new Annex VII C, which addresses some of the risks of chatbots highlighted in Section III.

Firstly, Article 28b incorporates some principles regarding data used in training. Article 28b(e) requires the providers of foundational models to "process and incorporate only datasets that are subject to appropriate data governance measures". The text of the proposal seems to leave it open to providers to set the exact type of data governance measures for foundation models. It only provides one example of such governance measures, notably "measures to examine the suitability of the data sources and possible biases and appropriate mitigation". In addition, Annex VII C also requires a "description of the data sources used in the development of the foundational model".[84] These proposals therefore bring together a requirement regarding the quality of the data ("data governance measures") and another one regarding the transparency of the datasets used in training. These two aspects together sketch a system whereby developers and providers

[79] See reports from the press: <https://www.linkedin.com/posts/luca-bertuzzi-186729130_aiactfinalfour-column21012024pdf-activity-7155091883872964608-L4Dn?utm_source=share&utm_medium=member_desktop>; L Bertuzzi, "EU's AI Act negotiations hit the brakes over foundation models" (*Euractiv*, 10 November 2023) <https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/>.

[80] EU AI Act – Parliament Amendments, Recital 60(e).

[81] EU AI Act – Parliament Amendments, Recitals 60(f) and (g).

[82] API stands for "application programming interface", a type of software interface that allows two systems to communicate. An API allows developers and users to access and fine-tune the underlying foundation model for their purposes without essentially modifying it. Examples of foundation models distributed via APIs are OpenAI's GPT-4 and Anthropic's Claude. See E Jones, "Explainer: what is a foundation model?" (*Ada Lovelace Institute*, 17 July 2023) <https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/>.

[83] EU AI Act – Parliament Amendments, Recital 60(f).

[84] EU AI Act – Parliament Amendments, Proposed modification to Annex VII C.

of foundation models are required to develop and follow their internal procedures to ensure the quality of the data, whereas the open character of the data sources should allow public scrutiny on the part of regulatory bodies, independent researchers, civil society and the press. Although this system seems to meet some of the requirements for keeping foundation models open,[85] recent research has shown that the currently available commercial chatbots are far from actually meeting these requirements.[86] In addition, the EU AI Act is still in the process of being negotiated, even though commercial chatbots have been deployed for almost a year. In the foreseeable future, this lack of regulation and accountability will continue, at least in the EU, further exacerbating the opacity of the training data and, arguably, the discriminatory or biased characters of the commonly used datasets.[87]

The EU AI Act – Parliament Amendments also requires the providers of foundation models used for generative AI, such as LLMs, to "document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law".[88] This provision needs to be read in conjunction with Article 4 of the Copyright Directive,[89] according to which training and the retention of data for training are allowed unless the holder of a right over content has expressly, including in a machine-readable format, excluded such use of their work. It seems, therefore, that the EU is consolidating its approach to allowing the use of copyrighted materials in the training of foundation models under the sole conditions of transparency and provided that the copyright holder has not opposed such use. This is in contrast with the present situation in the USA, where the issue is currently unsettled under copyright law and is the object of extensive litigation.[90] Although this approach allows machine learning and foundation models – which rely on large amounts of data – to exist and therefore may favour innovation, it also disregards the moral and economic rights of creators, as recently highlighted by authors and representatives of the creative industries.[91] And even if copyright may not ultimately be the panacea through which such moral and economic rights are guaranteed to single creators,[92] it seems necessary to reflect upon and devise legal solutions regarding how human creativity and content production can be preserved in the face of the rise of chatbots, most of which are profitable products of international commercial conglomerates.

Another issue that is addressed by the EU AI Act – Parliament Amendments is the labelling of machine-generated content. Nonetheless, the current text seems to require disclosure in the form of clear labelling or watermarking only for generated content that qualifies as a "deep fake",[93] defined as "content that would falsely appear to be authentic or truthful, and which features depictions of persons appearing to say or do things they did not say or do".[94] At this stage, therefore, there are no requirements in the EU AI Act – Parliament Amendments regarding disclosure that text has been machine generated in

[85] F Ferrari et al, "Foundation models and the privatization of public knowledge" (2023) 5 Nature Machine Intelligence 818, 819.

[86] R Bommasani et al, "Do Foundation Model Providers Comply with the EU AI Act?" (*Stanford Center for Research on Foundation Models, Institute for Human-Centered Artificial Intelligence*) <https://crfm.stanford.edu/2023/06/15/eu-ai-act.html>.

[87] On the phenomenon of the "naturalisation of a dataset", see supra, note 28, and accompanying text.

[88] Currently, this provision is Art 28(b)(4)(c) of the EU AI Act – Parliament Amendments.

[89] 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 and amending Directives 96/9/EC and 2001/29/EC (DSM Directive), PE/51/2019/REV/1, OJ L 130, 17.5.2019, pp 92–125.

[90] On these two aspects, see Lucchi, supra, note 52.

[91] See ibid, notably the various references in footnote 79.

[92] X Tang, "Copyright's Techno-Pessimist Creep" (2021) 90 Fordham Law Review 1151, 1185–87.

[93] EU AI Act – Parliament Amendments, Art 52(1).

[94] EU AI Act – Parliament Amendments, Art 3(1) Recital (44e).

general but only insofar as it constitutes a "deep fake".[95] This solution seems good overall considering that a blanket requirement to watermark text generated by AI may indeed take some of the utility out of these new tools, and even put certain individuals or groups at a disadvantage. For example, using a chatbot to proofread written communications could be a way for non-native speakers to perform certain tasks or access certain services from which they would otherwise be excluded.

Concerning the risks of harm and inaccuracies, the EU AI Act – Parliament Amendments lays down a series of compliance requirements aimed at making known information about the capabilities, performance, limitations and risks of foundation models, as well as the measures taken by the provider to mitigate such risks.[96] At the current stage, these obligations only require disclosure and do not seem to require any specific level of risk-proofing or action on the part of the provider of the LLM. Ample reference is made to possible future benchmarks and industry standards,[97] which the EU lawmaker expects will also be facilitated by newly established bodies.[98] In addition, these disclosure obligations apply to the providers of foundation models, whereas deployers of products based on such foundation models are not covered. Indeed, the idea behind this choice seems to be that entities wishing to use a foundation model, such as an LLM, to create and deploy a product should be able to use the public information on the foundation model to ensure compliance with the future European regulation.[99] In reality, in the current state of the market for chatbots, this exclusion may not be very relevant because the main commercial chatbots on the market are indeed those launched by the same companies having developed the underlying LLMs.[100] However, in the future, should commercial products be built and deployed on the market by a different entity than the one having developed and put on the market the LLM, such an exclusion might create regulatory loopholes and possibly a lack of accountability.

Another related aspect regards the moderation of machine-generated content. Article 28(4)(c) EU AI Act – Parliament Amendments requires providers of products such as text-generating LLMs to "train, and where applicable, design and develop the foundation model in such a way as to ensure adequate safeguards against the generation of content in breach of Union law in line with the generally-acknowledged state of the art, and without prejudice to fundamental rights, including the freedom of expression".[101] In the current state of generative chatbots, this requirement seems to place an overwhelming burden on providers of LLMs, in particular because LLMs may provide false information due to the way in which they produce text, which has no link to actual meaning but is merely based on a statistically probable combination of words. In this respect, could a false statement about the content of EU law be "in breach of EU law"? Conversely, would a statement that encourages discrimination be in breach of EU law because it is contrary to Article 21 of the EU Charter of Fundamental Rights? And, if the affirmative is the case, it should be for the provider to determine in advance the extent to which such a statement could be considered "not in breach of EU law" under the EU's view of freedom of expression. Although the concern underlying this provision seems understandable, the future compromise text should retain a wording that empowers providers rather than one that could only elicit further doubts and, possibly, litigation. Although not a silver bullet, a step

[95] Contradicting the conclusions of Bommasani et al, supra, note 86.

[96] EU AI Act – Parliament Amendments, Annex VIII – Section C, points (6) to (8).

[97] Art 28(b)(2), last paragraph. See also Annex VII C requiring providers of foundational models to provide "a description of the model's performance, including on public benchmarks or state of the art industry".

[98] EU AI Act – Parliament Amendments, Art 58(a).

[99] EU AI Act – Parliament Amendments, Recital (60g).

[100] See, for example, the table in Bommasani et al, supra, note 86.

[101] EU AI Act – Parliament Amendments, Section I.

in this direction could be to only refer to the EU Charter of Fundamental Rights as a benchmark of legality rather than the whole of EU law.

Finally, with respect to environmental concerns and, more broadly, to the computation utilised for the development of foundation models, the EU AI Act – Parliament Amendments requires providers of foundation models to disclose the model size, computer power and training time used.[102] It also requires providers to disclose the energy consumption of the model and take steps to make the training of foundational models more sustainable.[103] These requirements with respect to energy consumption are not only applicable to foundation models. The EP has added energy-saving requirements for AI systems in general.[104] This is an important step in leadership for the European legislator because providers and developers of foundation models will have to adapt to the stricter European standards to access the EU market and probably would not differentiate for other markets given the related costs. In this way, the EU legislation might foster a positive cycle regarding the achievement of more sustainable AI.

V. Conclusion

At the dawn of the commercialisation of chatbots leveraging LLM technology, and in view of their potentialities, the legal system is called to swiftly respond to the risks that they pose to individuals and society. New technologies are bringing about a new cognitive revolution[105] that will prompt humans to adapt to the new methods of information processing and communication that are brought about by AI-based technologies such as LLMs.

The role of law in this scenario is crucial. Technological inventions are not neutral, nor are they good per se. On the contrary, any new system embeds values, whether we like such an idea or not.[106] Accordingly, lawyers and policymakers should take a hard look at the potentialities and risks of chatbots leveraging LLMs and create a regulatory and legal framework that is able to steer this technology towards the common good and a future in which humans are empowered rather than overwhelmed by it.

Competing interests. The author declares none.

[102] EU AI Act – Parliament Amendments, Annex VII C.

[103] EU AI Act – Parliament Amendments, Art 28b.

[104] See, for example, Recitals 46, 46(a) and 46(b).

[105] RJ Neuwirth, *The EU Artificial Intelligence Act Regulating Subliminal AI Systems* (1st edition, London, Routledge 2022). See also the research into the effects of human–AI collaboration with respect to the game of Go: Shin et al, supra, note 51.

[106] H Nissenbaum, "How Computer Systems Embody Values" (2001) 34 Computer 118, 120.

Cite this article: S Migliorini (2024). ""More than Words": A Legal Approach to the Risks of Commercial Chatbots Powered by Generative Artificial Intelligence". *European Journal of Risk Regulation* **15**, 719–736. https://doi.org/10.1017/err.2024.4