Originally published in: Lindén, Krister/Niemi, Jyrki/Kontino, Thalassia (Eds.): Proceedings of the CLARIN Annual Conference 2023. 16 – 18 October 2023, Leuven, Belgium. – Utrecht: Clarin, 2023. Pp. 53-56. "

Do Chatbots Dream of Copyright? Copyright in AI-generated Language Data

Paweł Kamocki
IDS Mannheim
Germany
kamocki@ids-mannheim.de

Toby Bond
Bird & Bird
London, UK
toby.bond@twobirds.com

Krister Lindén
University of Helsinki
Finland
krister.linden@helsinki.fi

Thomas Margoni
KU Leuven
Belgium
thomas.margoni@kuleuven.be

Abstract

For language scientists, a *prima facie* advantage of AI-generated data over human-created content is that AI outputs are generally regarded as free from copyright. This submission addresses this issue in some detail.

1 Introduction

2023 is the year of a rabbit according to the lunar calendar, but in Europe it is most likely to be remembered as the year of Artificial Intelligence. It is safe to say that such events as the launch of Chat-GPT (in November 2022) or of GPT-4 have already revolutionized the way in which language data are generated. This revolution has not been unnoticed by the CLARIN community. The new perspective that AI opens up, is to create fully synthetic data according to the specifications of a researcher.

In branches of science where data for language modelling is scarce, or access to it is limited by (usually copyright or data protection) laws, protected, e.g., medical sciences, behavioral sciences, etc., the researchers can ask an AI model to generate new data for large categories, thereby avoiding the legal barriers. The model can also be used for creating more data for small categories to make the data more balanced and less biased. However, the bias reduction needs to be verified so that the additional data does more than just amplify the prejudice or bias in the original data.

For language scientists, a *prima facie* advantage of AI-generated data over human-created content is that, as it is generally agreed upon, AI outputs are not protected by copyright. This abstract addresses this issue in some detail.

The main reason for the absence of copyright in AI-generated data is their lack of human authorship (Section 2). However, the re-use of certain AI outputs may be in a legal grey area (Section 3). The introduction of a property right in AI outputs is seen by some as an answer to the challenges presented by the development of generative AI (Section 4); however, little evidence of this is found in the UK, where computer-generated works have been protected by a property right since 1988 (Section 5).

2 Lack of human authorship as an obstacle to copyright protection of AI outputs

The argument commonly used to refuse copyright protection of AI-generated content is lack of human authorship. Indeed, the use of the word 'author' (Cambridge Dictionary: 'a person who begins or creates something') seems to indicate that only works created by humans can be protected by copyright.

Although human authorship is not expressly required by the Berne Convention, this landmark international treaty uses words like 'nationality' (esp. Art. 3), 'honour' (Art. 6bis(1)) and 'death' (Art. 6bis(2), 7, 7bis) to refer to the author, which clearly points at a human being. The same conclusion can be drawn from the EU Directive 2006/116/EC on Copyright Term (see esp. Art. 1(4)).

It is true that many copyright systems accept 'corporate ownership' of copyright, where copyright is held *ab initio* by a legal person and not the human author. This is for example the case under the work for hire doctrine, where copyright in a work created by an employee belongs *ex lege* to the employer.

Even if one overlooks the absence of human authorship, AI-generated outputs could not be protected by copyright, as they would not satisfy the originality criterion, at least in the European Union. According to the CJEU, a work is original if it constitutes its author's own intellectual creation, i.e. it reflects the author's personality, which is the case when the author can express their creative abilities by making free and creative choices[1]. Conversely, when technical considerations, rules and constraints leave no room for free choices, there can be no originality and therefore no copyright protection[2]. Since, arguably, AI outputs do not reflect the personality of their authors, and their generation follows technical constraints, they also do not meet the originality criterion and are not eligible for copyright protection.

The US Copyright Office followed this approach already in 2018 when it refused to register a machine-generated image[3]. In practice, however, the legal status of AI-generated works is more complex than it may *appear prima facie*.

3 Grey areas related to AI outputs

AI does not (yet) generate outputs autonomously; the generative process is always initiated by a human who prompts the application with an idea in their mind. At least according to the dictionary definition, this human initiator can still be referred to as 'author' ('a person who *begins* or creates something'), even though the actual expression of the work (protectable by copyright, unlike the initial idea) is generated (or at least assisted) by AI.

Drawing a line between outputs with sufficient human involvement to 'deserve' copyright protection ('AI-assisted') and those without it ('AI-generated') is an extremely delicate task (cf. the 4-step test in Hugenholtz and Quintais, 2021), and courts' views on this issue are susceptible of evolving over time.

Such was the case with, e.g., photography, which was admitted in the realm of copyright several decades after the technology was popularized, and even today it is not recognized in the Berne convention as equal with other types of works (Art. 7(4) allows for a shorter term of protection for photographic works). In early decisions involving photographs[4] courts emphasized the role of the human photographer in, e.g., selecting the lighting, a task that is (or at least can be) fully automated in modern digital cameras, which does not seem to affect copyrightability of digital photographs (Margoni, 2014). AI outputs may follow the same trajectory, and the degree of human involvement required by courts for copyright protection may be gradually lowered. After all, since the beginning of time, almost all forms of human expression have employed some form of technology, be it very rudimentary.

In its recent policy statement, the US Copyright Office (2023) also opted for a somewhat nuanced approach to registering AI-generated works. In the Office's view, merely prompting a machine is not enough to claim authorship in the output (no matter how elaborated the prompt, according to the Office it only functions as an instruction to a commissioned artist). However, copyright can be claimed where AI outputs are arranged by a human in a creative manner, or modified to a degree that meets the threshold of creativity. This is illustrated by the Office's decision regarding a comic book "Zarya of the Dawn"[5], in which all images were generated by AI. The comic book as such (the plot, the texts) were deemed eligible for registration, although individual AI-generated images were excluded therefrom.

Another grey area regarding copyright in AI outputs is linked to their relationship with the input data used to train the underlying model. Although the use of copyright-protected content to train AI models is generally (under certain conditions) allowed under the exceptions for Text and Data Mining (Kamocki et al., 2018), the legal status of AI outputs is rather unclear. Carlini et al. (2021) have shown that certain text data AI models may sometimes 'regurgitate' portions of training material, which contributes to significant lack of legal certainty regarding copyright status of such outputs, especially considering that according to the CJEU excerpts as short as 11 consecutive words may be protected by copyright in certain circumstances. Even without regurgitating verbatim copies of training data, some (e.g., Gervais,

[1] See esp. CJEU *Infopaq* (C-5/08) and *Painer* (C-145/10)
[2] CJEU *Football Dataco* (C-604/10)
[3] https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf
[4] See esp. *Burrow-Giles Lithographic Co. v. Sarony*, 111 U.S. 53 (1884)
[5] https://www.copyright.gov/docs/zarya-of-the-dawn.pdf

2022) have argued that AI outputs are derivatives, derived from the training material, which would also impact their copyright status. This lack of legal certainty is illustrated by a recent US lawsuit, in which Getty Images sued Stability AI for allegedly using their images to train an AI model[6].

4 Towards (Property) Rights in AI Outputs?

In February 2023 it was reported that ChatGPT is listed as author or co-author of over 200 books available on Amazon (Nolan, 2023). One can only imagine the number of books and other texts that were 'secretly' generated by AI and passed as human creations. As purely AI-generated texts are generally in the public domain, they can fall victim to 'copyfraud', i.e. a false copyright claim (e.g. by simply signing an AI-generated text with one's name, as a pretended human author). In the current state of law, 'copyfraud', although certainly unethical, is usually not a punishable violation. In fact, the Berne Convention (Article 15(1)) and the EU Directive 2004/48/EC on the enforcement of IP rights (Article 5) establish a presumption of ownership for those whose name 'appear on the work in the usual manner'.

Already in the 1960s it was argued (Demsetz, 1967) that technological progress will necessarily be accompanied by the creation of new property rights, mostly to guarantee legal certainty of transactions and to prevent market failure. Indeed, in the last decades new property rights have been created, such as the *sui generis* database right, or the right in computer-generated works in the UK (see below).

Already in 2020 the European Parliament took the view that AI-outputs 'must' be protected under Intellectual Property Rights in order to encourage investment and improve legal certainty, and called the Commission to reform EU law accordingly. Such statements from the Parliament should, however, be regarded as devoid of any legal meaning. However, in a recent response[7], the Commission stated that 'the issue of AI-generated works does not deserve a specific legislative intervention'. Moreover, many European IP scholars criticize the idea of introducing new property rights (Bulayenko et. al, 2022).

On the other hand, in recent years the Commission was active in proposing governance-based (as opposed to property-based) regimes for data, including AI-generated data. This follows an attempt to introduce a data producers' right (Gangjee, 2022). These regimes, introduced e.g. by the Data Governance Act or the Data Act, are focused on rights of users, enabling access and portability of data (that companies want to keep 'secret'), rather than on recognizing monopolies (property rights) in the data (Margoni & Kretschmer, 2022). This can be a novel approach to regulating AI, both at the input end (e.g., by recognizing 'artist data', distinct from copyright in literary, artistic and scientific works), and at the output end.

For now, the re-use of AI outputs is mostly regulated by contracts, especially Terms and Conditions of related online services, which tend to vary significantly. For example, Terms of Use of ChatGPT allow for the generated content to be reused for any purposes, including commercial ones ('such as sale or publication'), with an important exception: the use of ChatGPT outputs to develop models that compete with OpenAI is prohibited. A similar prohibition can be found in Bard's Terms of Service. Bing's Terms of Use for its consumer-focused product only allow for the generated content to be reused 'for personal and non-commercial purposes'.

It should be noted here that if the outputs of these applications are not protected by copyright, copyright exceptions, including the TDM exceptions, cannot apply to them, and so the above mentioned Terms and Conditions cannot be overridden by such exceptions, as long as the contracts are enforceable.

Some language models, such as BERT or GPT-2, are also available under open source licences (Apache 2.0 and MIT, respectively), which impose no restrictions on the use of their outputs. However, more recent versions of GPT, starting from GPT-3, are publicly available only through a web API (i.e., subject to Terms and Conditions), and this trend is likely to continue with subsequent iterations of the most performant language models.

5 UK's Experience with Protection of Computer-generated Works

UK's Copyright, Designs and Patents Act of 1988 contains (since its adoption) a provision on computer-generated works (s9(3)). These works, defined as works 'generated by computer in circumstances such that there is no human author of the work', are protected by copyright (which, in the continental tradition,

[6] Getty Images (US), Inc. *v.* Stability AI, Inc. (1:23-cv-00135).

[7] https://www.europarl.europa.eu/doceo/document/E-9-2023-000479-ASW_EN.pdf (last access: 27.04.2023).

would be classified as a 'related' or 'neighbouring' right rather than copyright *stricto sensu*) for 50 years following their creation (s12(7)). The right belongs to 'the person by whom the arrangements necessary for the creation of the work are undertaken' (referred to as 'author'). Somewhat paradoxically, in order to qualify for protection, computer-generated works, like all other works, have to meet the criterion of originality (which historically was understood in the UK as involving a degree of 'labour, skill and judgement', but under the influence of the CJEU, a more author-centric approach to originality, presented above, was adopted). Similar provisions exist also in Ireland, New Zealand and South Africa.

Although it seems tempting to use this provision, adopted with the intention to regulate re-use of works such as satellite photographs, to AI-generated content, this has never been done by UK courts. In fact, case law involving this provision is extremely scarce, and the provision has been described as 'unclear and contradictory'. In a recent public consultation, the UK Intellectual Property Office listed computer-generated works as one of the issues to be addressed by the legislator. In its 2022 response, however, the government stated that, as there is no evidence that the provision is harmful, and 'any changes could have unintended consequences', especially given that the development of AI is still in its early stages. In the same statement, the government also declared that they will keep the provision under review and may remove, replace or amend it if the evidence supports this[8].

References

Bulayenko, O., Quintais, P. J., Gervais, D. & Poort, J. (2022). *AI Music Outputs: Challenges to the Copyright Legal Framework*. ReCreating Europe Report. https://doi.org/10.5281/zenodo.6405796

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A. & Raffel, C. (2021). Extracting Training Data from Large Language Models. *arXiv: 2012.07805.* https://doi.org/10.48550/arXiv.2012.07805

Demsetz, H. (1967). Toward a Theory of Property Rights. *The American Economic Review*, 57, 2, 347-359.

European Parliament. (2020). *Resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies* (2020/2015(INI))

Gervais, D. J. (2022). AI Derivatives: the Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines. *Seton Hall Law Review*, 53, 1111-1136. http://dx.doi.org/10.2139/ssrn.4022665.

Gangjee, D. S. (2022). The Data Producer's Right: An Instructive Obituary. [in:] Lim, E. & Morgan, P. (eds), *The Cambridge Handbook of Private Law and Artificial Intelligence*, Cambridge University Press.

Hugenholtz, P.B., & Quintais, J.P. (2021). Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output? *International Review of Intellectual Property and Competition Law*, 52, 1190–1216. https://doi.org/10.1007/s40319-021-01115-0

Kamocki, P., Ketzan, E., Wildgans, J. & Witt, A. (2018). New exceptions for Text and Data Mining and their possible impact on the CLARIN infrastructure. *Selected papers from the CLARIN Annual Conference 2018*

Margoni, T. (2014). The Digitisation of Cultural Heritage: Originality, Derivative Works and (Non) Original Photographs (December 3, 2014). http://dx.doi.org/10.2139/ssrn.2573104

Margoni, T. & Kretschmer, M. (2022). A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. *GRUR International*, 71(8), 685–701. http://dx.doi.org/10.2139/ssrn.3886695

Nolan, B. (2023). More than 200 books in Amazon's bookstore have ChatGPT listed as an author or coauthor. *Business Insider*, February 23, 2023. https://www.businessinsider.com/chatgpt-ai-write-author-200-books-amazon-2023

US Copyright Office. (2023). *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence*. 16190 Federal Register, vol. 88, no. 51, 37 CFR Part 202.

[8] https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/artificial-intelligence-and-intellectual-property-copyright-and-patents (last access: 27.04.2023).