Mind the Ownership Gap?
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

Aleksei Kelli
University of Tartu
Estonia
aleksei.kelli@ut.ee

Andrius Puksas
Vytautas Magnus University
Lithuania
andrius.puksas@vdu.lt

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 contribution addresses this issue in some detail.

1 Introduction

2023 was the year of the rabbit according to the lunar calendar, but in Europe it will likely 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 modeling is scarce, or access is limited by confidentiality or (usually copyright or data protection) laws, e.g., medical or behavioral sciences, the researchers can ask an AI model to generate new synthetic 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.

Examples of synthetic data use in commercial research include Amazon using synthetic data to train Alexa's language system with available sample utterances as templates generating new data by combining and varying the templates. Google's Waymo uses synthetic visual data to extend its training data for self-driving cars with more complex but infrequent scenarios virtually adding more agents for the AI to cope with. American Express and J.P. Morgan generate statistically accurate synthetic data from financial transactions for more sophisticated fraud detection, and Roche uses validated synthetic medical data as a replacement for clinical research data to develop AI healthcare algorithms with massive amounts of personal health data, while minimizing privacy concerns[1].

[1] Types of synthetic data and 4 real-life examples (2022): https://www.statice.ai/post/types-synthetic-data-examples-real-life-examples (last access: 13.02.2024)

For 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 contribution addresses this issue in some detail.

The main reasons for the absence of copyright protection for AI-generated data is their lack of human authorship (Section 2), as well as – closely related – lack of originality (Section 3). However, the re-use of certain AI outputs may be in a legal grey area (Section 4). 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 5), despite the fact that little evidence of this is found in the UK, where computer-generated works have been protected by a property right since 1988 (Section 6).

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. The author is indeed placed at the very heart of modern copyright law, which was largely modeled after the French tradition of *droit d'auteur*, or *author's right*. The crucial role of the author in copyright law is illustrated by the fact that the author is the default holder of both economic and moral rights; moreover, the term of protection is also linked (at least in most cases) to the death of said author.

According to the general dictionary definition of the word "author", only human beings seem to be able to qualify as such. For example, the Cambridge Dictionary defines "author" as a 'a person who begins or creates something'; other dictionary definitions also seem to reserve this status to humans. But is this also the case in legal context? The question is worth asking, as legal texts often attribute specific meaning to everyday words.

The Berne Convention does not define "author", and does not expressly require human authorship for copyright-protected works. Nevertheless, upon closer inspection, this landmark international treaty is clearly based on the assumption that the author is a human being. For example, according to Article 3(2) "Authors who are not *nationals* of one of the countries of the Union but who have their *habitual residence* in one of them shall, for the purposes of this Convention, be assimilated to nationals of that country" (italics added by the authors). Moreover, Article 6bis provides that "the author shall have the right to (...) object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his *honor* or *reputation*", and adds that this right "shall, after [the author's] *death*, be maintained (...)". Finally, Article 7 defines the term of copyright protection as "*life* of the author and fifty years after his *death*", and Article 7bis further specifies that this term should be calculated "from the *death* of the last *surviving* author". Therefore, it appears clearly that under the Berne Convention only humans can be authors, as, unlike AI, they are mortal, have a nationality and a place of residence, as well as honor and reputation.

In EU law, the same conclusion can be drawn from the Copyright Term Directive (2006/116/EC), whose Articles 1 and 2 also refer to the author's death while defining the duration of copyright protection (which under EU law is longer than required by the Berne Convention, i.e. seventy years after the death of the author). Most, if not all, national laws[2] also contain similar provisions, tying copyright terms to the death of the author. Moreover, German Copyright Act (*Urheberrechtgesetz*) defines (in its §7) "author" as "the creator of a work", which also seems to reserve this status to humans.

Recently, the UK Supreme Court[3] ruled that Artificial Intelligence cannot be regarded as an "inventor" under patent law, and that only human beings can be "inventors". This, of course, is not directly related to copyright, but one can expect that if AI cannot be an "inventor", a fortiori it cannot be an "author".

In the US, the US Copyright Office (2023) also recognises that "copyright can protect only material that is the product of human creativity". In recent years, the Office has refused to register AI-generated images on the grounds of lack of human authorship: this was the case of a rather appealing image entitled *A recent entrance to paradise* (US Copyright Office (2022); *Fig. 1*) – the decision was later upheld by

[2] E.g., Article L. 123-1 of the French Intellectual Property Code, §64 of the German Copyright Act, §302 of the US Copyright Act (17 USC), s12 of the UK's Copyright, Designs and Patents Act, etc.

[3] Thaler (Appellant) v Comptroller-General of Patents, Designs and Trade Marks (Respondent) [2023] UKSC 49.

a District Court[4], who also found the image uncopyrightable and "absent any human involvement" – as well as a prize-winning image *Théâtre d'opéra spatial* (US Copyright Office (2023b); *Fig. 2*). The latter case is particularly interesting: the Copyright Review Board of the US Copyright Office found that the image, which for the most part was generated by Midjourney, lacked human authorship, and the applicant's input in generating it ('at least' 624 text prompts) was not sufficient to make him the author. Despite the fact that the modifications made by the applicant in Adobe Photoshop, if considered in isolation, might have qualified for copyright protection, the Midjourney-generated basis could not, and therefore the final image could not be registered as copyright-protected.

Figure 1: A recent entrance to paradise *(AI-generated)*

[4] US District Court of Columbia, 18.08.2023, Case 1:22-cv-01564-BAH, https://storage.courtlistener.com/recap/gov.uscourts.dcd.243956/gov.uscourts.dcd.243956.24.0_2.pdf (last access: 13.02.2024)

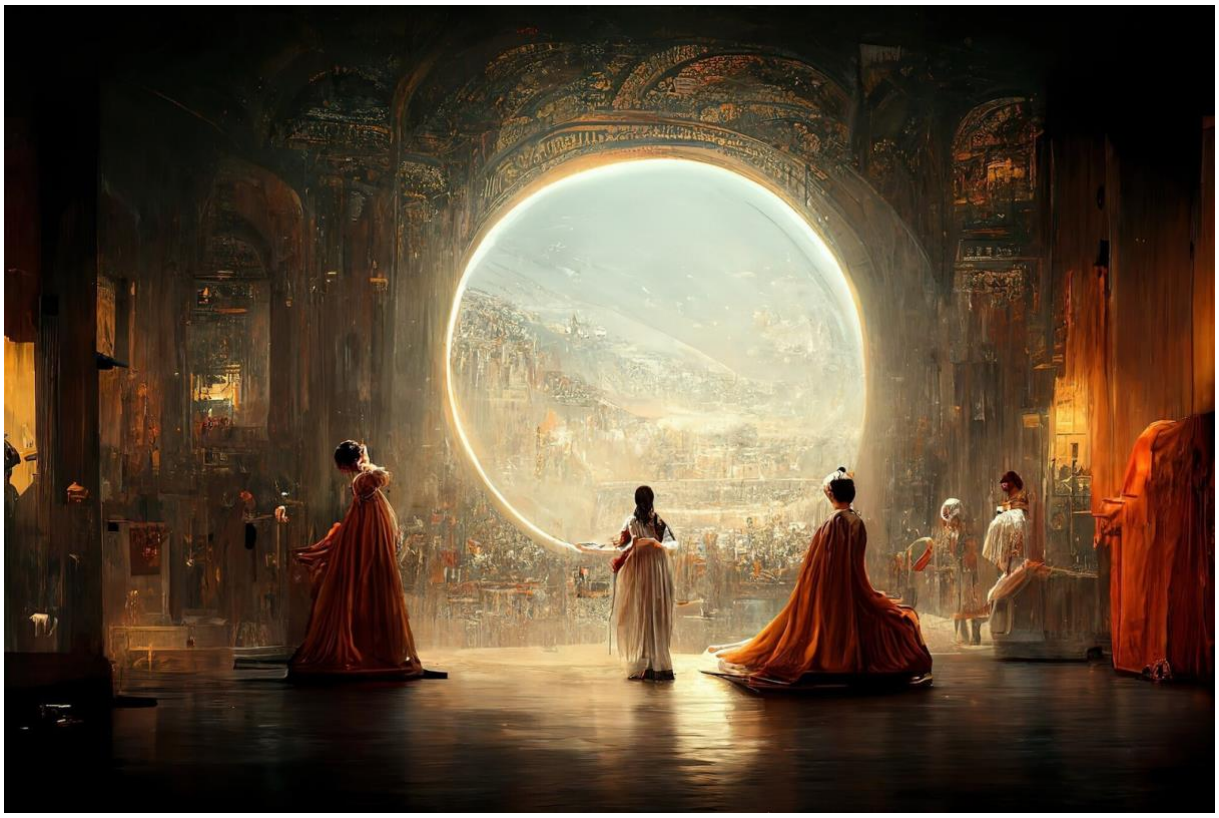

Figure 2: Théâtre d'opéra spatial *(AI-generated with human-made adjustments)*

At the same time, many copyright systems accept 'corporate ownership' of copyright, i.e. a situation where copyright is held *ab initio* by a legal person (a company, an employer) and not the human author. This is for example the case under the traditional anglo-saxon doctrine of *work for hire*, where copyright in a work created by an employee belongs *ex lege* to the employer[5]. Initial ownership of copyright by a corporation is therefore a well-established solution, which some praise as pragmatic and promoting investment (or even innovation). In fact, even in such an author-oriented copyright system as the French *droit d'auteur*, economic rights in a collective work (*oeuvre collective*) the creation of which is initiated and supervised by a legal person who then disseminates the work under its name[6] belong *ab inito* to the legal person, and not the actual human authors. In the field of software, Article 2 of the Directive 2009/24/EC on computer programs attributes the economic rights in software created by employees in the execution of their duties to the employer. The Article goes as far as to admit (in paragraph 1) that, where legislation of a Member State allows it, a legal person can be considered author of a computer program.

It appears, therefore, that many national laws, and even, to an extent, EU law, can tolerate a situation where initial ownership of copyright is attributed to a legal, and not a natural person. This, however, is not enough to solve the issue of AI-generated works, since AI in itself obviously has no legal personality, and attributing copyright to the company that provides a generative AI tool (such as Open AI, the provider of Chat GPT) would be a dubious solution to say the least.

3 Lack of originality as an obstacle to copyright protection of AI outputs

Another theoretical obstacle on the path to copyright protection that AI-generated works would have to face is the originality requirement.

Originality (in some copyright traditions, e.g. in Germany and in Poland, also referred to as "individuality") is the main condition for copyright protection. At the same time, it is a very elusive

[5] Cf., for example the definition of a "work made for hire" in § 101 of the US Copyright Act 1976, or s11(2) of the UK's Copyright, Designs and Patents Act of 1988.
[6] Cf. Article L. 113-2 para 3 of the French Intellectual Property Code.

concept, which for a long time was escaping any efforts toward international harmonisation. Briefly put, two approaches to originality can be distinguished: a subjective one, which emphasises the relation between the work and its author (a work is original if it carries a "personal mark" of the author) and an objective one, which focuses on elements such as skilled effort invested in the creation and novelty (absence of copy) of the resulting work.

Since CJEU's 2009 landmark decision in the *Infopaq* case[7], the EU subscribes to the subjective approach, even though it contradicts long-standing traditions of some Member States' national copyright laws. In the *Infopaq* case, the CJEU applied the definition of originality as "author's own intellectual creation", which was already present in EU law, to all copyright-protected works. Incidentally, this definition is also very close to the traditional German concept of "personal intellectual creation" (*persönliche geistige Schöpfung*)[8]. This was further elaborated in subsequent CJEU's decisions; most notably in *Painer*[9], where the Court ruled that the originality requirement is met "if the author was able to express his creative abilities in the production of the work by making free and creative choices". By making such choices, the author "stamps the work with his personal touch", so that the work "reflects his personality". At the same time, the CJEU formulated various "negative conditions" for originality, i.e. conditions that, if met, prevent copyright protection; these include situations where the expression of the work is dictated by technical considerations[10], or other rules that leave no room for creativity[11]. Moreover, the CJEU also clearly stated (in *Football Dataco Ltd*)[12] that labour and skill alone are not enough to justify copyright protection of the outcome.

It seems that autonomous AI outputs cannot meet the originality criterion as defined by the CJEU, as generative AI tools may not allow the user to make free and creative choices during the creative process, and to leave his or her "personal touch" in the work. This also seems to be the position of Advocate General Trstenjak, who in her opinion in the *Painer* case stated that only human creations can be original (in the sense of being their author's own intellectual creations) and therefore qualify for copyright protection[13]. Moreover, one could expect that mere "skill and labour" invested by the user in prompting the generative AI tool are not enough to confer originality to the output.

US copyright law has a somewhat lower (or at least: more objective) standard of originality. In order to qualify as original under US law, a work has to be independently created by the author (i.e., simply, not copied from another work) and possess a minimal degree (modicum) of creativity, a "creative spark", "no matter how crude, humble or obvious it might be" (US Copyright Office, 2021). Arguably, at least some AI-generated works may pass this test. The US Copyright Office, however, systematically refuses to register AI-generated outputs not because of their lack of originality, but because of their lack of human authorship (cf. above).

4 Grey areas related to copyright protection of AI outputs

Lacking both human authorship and (subjective) originality, AI outputs may seem safely beyond the scope of copyright protection. However, this statement is not uncontroversial, and there are circumstances where AI outputs may be argued to meet the requirements for protection.

Firstly, 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. The main obstacle to copyrightability of AI outputs may therefore lie not in the law, but in the way our culture perceives authorship – and this can evolve over time, like it did in the past (Compagno, 2012).

For decades now, copyright theorists have been distinguishing between machine-assisted and machine-generated outputs. While machine-generated works are not protected by copyright (for the reasons discussed above), machine- (computer-, AI-) assisted works are characterised by a sufficient

[7] CJEU, C-5/08, 16.07.2009 (*Infopaq*).
[8] §2(2) of the German Copyright Act.
[9] CJEU, C-145/10, 1.12.2011 (*Painer*).
[10] CJEU, C 393/09, 22.12.2010 (Bezpečnostní softwarová asociace).
[11] CJEU, joined cases C-403/08 and C-429/08, 4.10.2011 (*Football Association Premier League Ltd*)
[12] CJEU, C-604/10, 1.03.2012 (*Football Dataco Ltd.*)
[13] Opinion of Advocate General Trstenjak delivered on 12 April 2011 in Case C-145/10, para 121

degree of human intervention to qualify for copyright protection. A vast majority of works are in some way assisted by a machine, including this very article, whose creation involved modern text processing software with, among other features, an in-built automatic spellchecker. This, however, does not change the fact that the article is, by any standard, protectable by copyright.

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 changing 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[14] 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 (2023a) also opted for a somewhat nuanced approach to registering AI-generated works. The key criterion seems to be whether the "traditional elements of authorship (literary, artistic, or musical expression or elements of selection, arrangement, etc.)" were "conceived and executed" by a man (assisted or not by a machine) or by a machine. In the Office's view (see above), merely prompting a machine is not enough to claim authorship in the output (no matter how elaborated or numerous the prompts, according to the Office they only function as "instructions to a commissioned artist", and the "traditional elements of authorship" are still executed by a machine). However, copyright can be claimed in cases where AI outputs are arranged by a human in a creative manner, or modified to a degree that meets the threshold of creativity.

This position was illustrated by the Office's recent decision regarding a comic book *Zarya of the Dawn* (US Copyright Office (2023c); Fig. 3) in which all images were generated by AI. The comic book as such (the plot, the dialogues) were deemed eligible for registration, although individual AI-generated images were excluded therefrom. However, the policy statement may seem inconsistent with the Office's decision concerning the image *Théâtre D'opéra Spatial* (Fig. 2, see above), which was also denied copyright protection. Although the image generated by Midjourney had been modified by the user, and the adjustments made might have been copyrightable on their own, the final result as submitted to the Office was not deemed eligible for copyright protection (Roose, 2022).

[14] See esp. Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53 (1884)

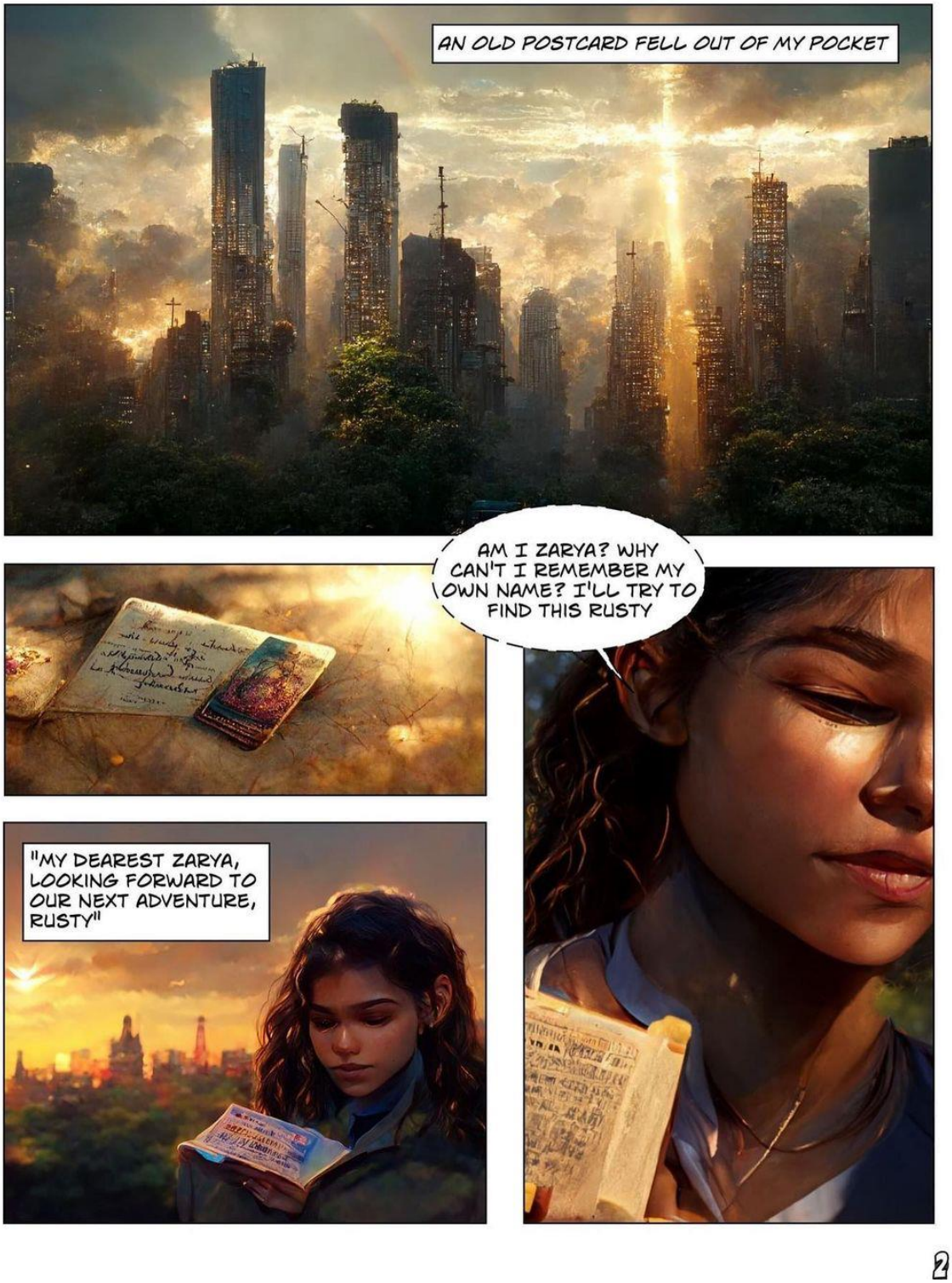

Figure 3: Panels from Zarya of the Dawn *(AI-generated images, texts by Kashtanova)*

Secondly, AI tools do not create outputs *ex nihilo*. Therefore, another gray area regarding copyright in AI outputs is linked to the relationship of these outputs with the data that were used to train the underlying model. Although in EU law the use of copyright-protected content to train AI models seems generally (under certain conditions) allowed under the exceptions for Text and Data Mining (Kamocki et al., 2018; Kelli et al., 2022), the copyright status of AI outputs remains rather unclear. Carlini et al. (2021) have shown that under certain conditions "training data extraction attacks" can be performed on GPT-2 which result in the model outputting text which resembles training material. The existence of

these techniques contributes to a lack of legal certainty regarding the copyright status of such outputs, especially considering that according to the CJEU excerpts as short as 11 consecutive words may be protected by copyright (Infopaq, C-5/08).

Even without regurgitating verbatim copies of training data, some (e.g., Gervais, 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 recent US lawsuits; e.g., Getty Images sued Stability AI for allegedly using their images to train an AI model[15], and NY Times sued Open AI and Microsoft for allegedly using their articles for this purpose (O'Brien, 2024). A group of 17 authors, including John Grisham and George R. R. Martin, went as far as to sue Open AI for "systematic theft on a mass scale" (Italie, 2023). As a matter of fact, there are a number of other lawsuits brought by authors against AI companies (see, Setty, 2023) or content producers accused of using AI techniques (Khalid, 2024). According to recent media reports, a Chinese court found a provider of an AI text-to-image tool guilty of copyright infringement; the tool (when prompted accordingly) generated images of Ultraman, a popular cartoon character, that were substantially similar to the original artwork (Costigan, 2024).

The opinion according to which AI-generated outputs are in fact infringing copyright in the data used to train the underlying model remains to be tested by European and US courts, and rightholders have so far struggled to consistently identify outputs which bear a resemblance to items of training data without wilfully contriving circumstances intended to create such resemblance[16].

Finally, it has also happened that, for fear of a successful copyright infringement lawsuit, platforms removed AI-generated content when pressured by rightholders (Snapes, 2023). Such content was, therefore, assumed to infringe copyright.

5 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 fact, the Berne Convention (Article 15(1)) provides that 'in order that the author of a literary or artistic work protected by this Convention shall, in the absence of proof to the contrary, be regarded as such, and consequently be entitled to institute infringement proceedings in the countries of the Union, it shall be sufficient for his name to appear on the work in the usual manner. This (…) shall be applicable even if this name is a pseudonym, where the pseudonym adopted by the author leaves no doubt as to his identity'. The same principle is repeated in the EU Directive 2004/48/EC on the enforcement of IP rights (Article 5). Both legal instruments establish a presumption of ownership for those whose name 'appear on the work in the usual manner'.

It seems, therefore, that it is enough for a user of a generative AI tool to sign his or her name on the output in order to benefit from a strong presumption of authorship, and become *de facto* enabled to sue others for copyright infringement. In this context, the act of signing automatically-generated content with one's name may appear controversial from the ethical standpoint, and it does constitute an act of copyfraud, but is not effectively punishable in the current state of the law.

One way out of this conundrum is the introduction of a transparency obligation, according to which all AI outputs would have to be clearly labeled as such. The proposed AI Act (European Commission, 2021) aims at addressing this issue in its Article 52, which on the one hand requires the providers of AI systems to design those systems in such a way as to inform users that they are interacting with AI, and on the other hand, obliges the users of image-, audio- or video-generating AI systems to disclose that the content resembling existing persons had been artificially generated. In the proposal, however (unlike in the French reform proposal discussed below), this obligation does not apply to AI-generated texts. Considering that such texts are practically indistinguishable from human-written ones (Casal & Kessler, 2023), such a requirement would meet serious evidence-related obstacles. It is conceivable to make the providers, and not the users, responsible for ensuring transparency of AI-generated text, e.g. by an

[15] Getty Images (US), Inc. *v.* Stability AI, Inc. (1:23-cv-00135).
[16] Cf. Open AI's rebuttal of NY Times' accusations: https://openai.com/blog/openai-and-journalism (last access: 13.02.2024)

imposed implementation of watermarking techniques. In the current state of the art, however, the feasibility of watermarking, particularly in shorter texts, seems doubtful.

If "enclosing" AI-generated content with a transparency obligation proves unworkable, other solutions to fill the perceived "void of ownership" (US Copyright Office, 2023b) would be to either extend the scope of copyright to include such works, or to create a new (property?) right to protect them.

Rather surprisingly, the authors of a proposal recently submitted to the French legislator opted for extending the scope of copyright. A copyright reform proposed in September 2023 (Assemblée Nationale, 2023) aims at introducing a series of rather revolutionary measures to protect the interests of creators against the influx of AI-generated creations. Firstly, this would include an express provision according to which rightholders' permission would be necessary to integrate a copyright-protected work "in an AI system", which at least *prima facie*, contradicts TDM exceptions (a part of EU *acquis*). Secondly, another new provision would state that copyright in content generated by AI without direct human intervention should belong to the authors of works that "enabled" the generation of the content. This not only contradicts the general lack of copyright in AI-generated works (due to lack of originality and human authorship), but is also extremely difficult to apply in practice, as it is rarely possible to determine a limited group of authors whose works "enabled" the AI system to generate a specific output; furthermore this may be regarded as a violation of EU law, as according to the CJEU originality (understood as "the author's own intellectual creation" – see above) is the only condition ("necessary and sufficient") for copyright protection[17]. Thirdly, the proposal also includes a system of collective rights management for AI-generated works. A designated collective rights management organisation would represent holders of rights in AI-generated works (i.e., according to the proposal, authors of works that AI used to generate content), perceive remuneration on their behalf and redistribute it among them. Fourthly, the proposal also aims at including a transparency obligation: all AI-generated outputs would not only have to be labeled as such, but also carry the names of all the authors that have enabled their creation (and who, therefore, would hold copyright in the work, as per the proposal). Especially in the case of longer AI-generated texts, this obligation is rather impossible to meet, because, among other reasons, it is often difficult to determine authorship in data obtained via web crawling. However, the French proposal also anticipates a situation where it is impossible to determine the *origin of works* (which, we believe, should be interpreted as authorship) that were used by an AI system; in such cases a levy (tax) would have to be paid to the abovementioned designated collective management organisation.

Although the proposed system of collective rights management and levies for AI-generated works may seem both controversial and impracticable, it has advocates among Europe's most renowned Intellectual Property scholars. In his recent article Senftleben (2023) argues that an output-oriented levy system, in contrast to remuneration for AI training activities, "does not weaken the position of the European AI sector and the attractiveness of the EU as a region for AI development. Even more importantly – Senftleben continues – an output-oriented AI levy system can be combined with mandatory collective rights management".

The creation of an entirely new exclusive right in AI outputs would be another possibility. As early as 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[18], 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

[17] CJEU, C-683/17, 12.09.2019 (*Cofemel*), para 30.

[18] https://www.europarl.europa.eu/doceo/document/E-9-2023-000479-ASW_EN.pdf (last access: 13.02.2024).

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.

6 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, 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[19].

7 Conclusion

AI-generated outputs which do not involve human creative input should, in principle, remain copyright-free, as they cannot meet the traditional criteria for copyright protection. However, AI-generated outputs may bear strong similarities to copyright-protected works, which causes significant tensions between the interests of authors, generative AI tools providers, and users of such tools. These tensions resulted in a series of lawsuits, and in the coming years some landmark court decisions are expected both in Europe and in the US.

Meanwhile, copyright scholars and legislators are pondering the possibility of extending the scope of copyright, or even introducing a new related right, to balance the interests at stake. The result of these

[19] https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/artificial-intelligence-and-intellectual-property-copyright-and-patents (last access: 13.02.2024).

debates should not, or at least not before several years, affect the possibility for scholars to use synthetically generated data for their research.

However, one should not lose sight of the fact that generative AI tools are generally available via web APIs, governed by Terms and Conditions, which are likely to regulate the way the tool can be used and the allowed uses of the outputs; since these outputs are not protected by copyright, copyright exceptions (e.g., for research or TDM) do not apply.

We do live in interesting times, certainly for copyright scholars.

References

Assemblée Nationale (2023), Proposition de loi visant à encadrer l'intelligence artificielle par le droit d'auteur, No. 1630, https://www.assemblee-nationale.fr/dyn/16/textes/l16b1630_proposition-loi.

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

Casal, J. E. & Kessler, M (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing, *Research Methods in Applied Linguistics*, Volume 2, Issue 2023, https://doi.org/10.1016/j.rmal.2023.100068.

Compagno, D. (2012). Theories of Authorship and Intention in the Twentieth Century: An Overview. *Journal of Early Modern Studies*, 2012, 1 (1), pp.37-53. hal-01846362

Costigan, J. (2024). China Rules AI Firm Committed Copyright Infringement. *Forbes*, February 29, 2024. https://www.forbes.com/sites/johannacostigan/2024/02/29/china-rules-ai-firm-committed-copyright-infringement/

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))

European Commission (2021). *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (AI Act) and Amending Certain Union Legislative Acts*. COM(2021) 206 final.

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

Italie, H. (2023). 'Game of Thrones' creator and other authors sue ChatGPT-maker OpenAI for copyright infringement. *Associated Press News*. https://apnews.com/article/openai-lawsuit-authors-grisham-george-rr-martin-37f9073ab67ab25b7e6b2975b2a63bfe

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*

Kelli, A., Tavast, A., Lindén, K. (2022). Building a Chatbot: Challenges under Copyright and Data Protection Law. In: Martin Ebers, Cristina Poncibò, Mimi Zou (Ed.). *Contracting and Contract Law in the Age of Artificial Intelligence*. (115−134). Hart Publishing. http://dx.doi.org/10.5040/9781509950713.ch-007

Khalid, A. (2024). Amazon's Road House reboot is accused of copyright infringement — and AI voice cloning. *The Verge*, February 28, 2024. https://www.theverge.com/2024/2/27/24085264/amazon-road-house-reboot-lawsuit-ai-cloning-copyright-infringement

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

O'Brien, M. (2024). ChatGPT-maker braces for fight with New York Times and authors on 'fair use' of copyrighted works, Associated Press News, https://apnews.com/article/openai-new-york-times-chatgpt-lawsuit-grisham-nyt-69f78c404ace42c0070fdfb9dd4caeb7.

Roose, K. (2022). An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

Senftleben, Martin, Generative AI and Author Remuneration (2023). International Review of Intellectual Property and Competition Law (IIC) 54 (2023), Forthcoming, Available at SSRN: https://ssrn.com/abstract=4478370 or http://dx.doi.org/10.2139/ssrn.4478370

Setty, R. (2023). Sarah Silverman, Authors Hit OpenAI, Meta With Copyright Suits. https://news.bloomberglaw.com/ip-law/sarah-silverman-authors-hit-openai-meta-with-copyright-suits

Snapes, L. (2023). AI song featuring fake Drake and Weeknd vocals pulled from streaming services. The Guardian. https://www.theguardian.com/music/2023/apr/18/ai-song-featuring-fake-drake-and-weeknd-vocals-pulled-from-streaming-services

US Copyright Office (2021). Copyrightable Authorship: What Can Be Registered, https://www.copyright.gov/comp3/chap300/ch300-copyrightable-authorship.pdf.

US Copyright Office, Copyright Review Board (2022). Second Request for Reconsideration for Refusal to Register A Recent Entrance to Paradise. https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf

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

US Copyright Office, Copyright Review Board (2023b). Second Request for Reconsideration for Refusal to Register Théâtre D'opéra Spatial. https://copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf

US Copyright Office (2023c). Zarya of the Dawn (Registration # VAu001480196). https://www.copyright.gov/docs/zarya-of-the-dawn.pdf