A FEW PRELIMINARY CONSIDERATIONS ON INFRINGEMENT OF COPYRIGHT BY ARTIFICIAL INTELLIGENCE

Paul-George BUTA[*]

Abstract

The issue of copyright infringement by artificial intelligence (AI) has become more salient, with the worldwide popularity of „Heart on My Sleeve" and with two other class action claims having been filed. The current piece aims to set out the core concepts, to list the possible issues to be addressed by further research and to suggest some perspectives to so address this matter. The most relevant from a copyright infringement perspective seem to be: (1) the distinction between inputs and outputs, i.e., the data the model is trained on and the product of the model; and (2) the existence of different models, each with a different level of human involvement and the existence, within each such model, of component algorithms, each with a different level of human input and relative importance of such. Taking all this into account clarifies the issues that could be incidental to an analysis of possible risks of copyright infringement by artificial intelligence models and also provides some guidance as to the perspectives from which such analysis could be conducted.

Keywords: *artificial intelligence, copyright, infringement, originality, machine learning.*

1. Introduction

Not longer than a few weeks ago there were numerous reports[1] in the media about a song named „Heart on My Sleeve" which had been removed from Spotify, TikTok, Apple Music and YouTube after amassing millions of streams because it was deemed to infringe the rights of artists Drake and The Weeknd, the song having been created with the help of artificial intelligence – or „AI" - (at least) in what concerns the vocal performance (which was wrongly attributed to the two aforementioned artists).

While the possible tensions between ever-more capable computers and copyright law have been espoused in the literature for more than 50 years,[2] most perspectives have been focused on whether works generated by 'creative computers' (or AI) should benefit from copyright protection (and, if so, who the rightsholder should be) and less so on the issues surrounding the infringement of existing copyright in human works and how such should be addressed.

The issue of infringement of copyright by AI has now apparently become more salient, with the worldwide popularity of „Heart on My Sleeve" and with two other class action claims having been filed. The first of these was filed on behalf of owners of copyrights in computer code which they had made available on the platform GitHub and is directed against GitHub, Microsoft (GitHub's owner) and OpenAI, the alleged infringement concerning the code posted by GitHub's users on the platform and its subsequent use by an AI product called Copilot (which is based on an AI called Codex which is used to convert natural language into code.[3] The second was filed by and on behalf of owners of copyright in photographs and is directed against Stability AI (the company offering the image-generation AI Stability Diffusion), Midjourney (who offers the image-generation AI of the same name) and DeviantArt (who offers DreamUp, a product relying on Stable Diffusion to produce images based

[*] Lecturer, PhD, Faculty of Law, „Nicolae Titulescu" University of Bucharest (e-mail: paul.buta@univnt.ro).

[1] See, *e.g.*, M. Sato, *Drake's AI clone is here — and Drake might not be able to stop him*, in The Verge, 1 May 2023, available at *https://www.theverge.com/2023/5/1/23703087/ai-drake-the-weeknd-music-copyright-legal-battle-right-of-publicity*, last time consulted on 06.05.2023; J. Coscarelli, *An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World*, in The New York Times, 19 April 2023, available at *https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html*, last time consulted on 06.05.2023.

[2] See, *e.g.*, K. F. Milde, Jr. *Can a Computer Be an „Author" or an „Inventor"?*, in Journal of the Patent Office Society no. 51 (1969), p. 378; T. L. Butler, *Can a Computer Be an Author - Copyright Aspects of Artificial Intelligence*, in Hastings Communication and Entertainment Law Journal no. 4 (1982), p. 707; P. Samuelson, *Allocating Ownership Rights in Computer-Generated Works*, in University of Pittsburgh Law Review no. 47 (1986), p. 1185; E. H. Farr, *Copyrightability of Computer-Created Works*, in Rutgers Computer and Technology Law Journal, no. 15 (1989), p. 63, all referenced in J. Grimmelmann, *There's no Such Thing as a Computer-Authored Work*, in Columbia Journal of the Law & the Arts no. 39 (2016), pp. 403-404, note 3.

[3] *Doe v. GitHub Inc.*, U.S. District Court for the Northern District of California, no. 3:22-cv-06823, referenced in J. Vincent, *The lawsuit that could rewrite the rules of AI copyright* in The Verge, 08.11.2022, available at *https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data*, last time consulted on 06.05.2023.

on text prompts).[4]

Thus, further analysis of the impact of AI 'generative' activities on the rules governing copyright infringement has become necessary. The current piece therefore aims to set out the core concepts, to list the possible issues to be addressed by further research and to suggest some perspectives to so address this matter.

We do not aim for a comprehensive analysis nor do we hope to arrive at definitive solutions. The field of AI is in a state of fervent development, with little transparency as to how the algorithms are developed and trained. The ever-increasing impact and the enlarging scope of the so-called AI 'generative' products make this analysis more salient by the day. Therefore this paper will hopefully have some use in the guiding of further research and in the grounding of such in some fundamental principles which should help focus future research so as to provide more consistency of approach.

2. The core concepts

Obviously, the most important concept relevant for our discussion is that of „artificial intelligence". The origins of the concept of AI are unclear, with Encyclopedia Britannica referring back to a 1947 Alan Turing conference in London where he had mentioned a „machine that can learn from experience" which would be made possible by the same machine being allowed to change its own instructions.[5]

John McCarthy is credited with coining the term „artificial intelligence" in 1956 when he invited researchers to a summer workshop on „artificial intelligence,"[6] him being convinced that any feature of human learning or intelligence can be described so precisely that a machine could be programmed to simulate it.[7]

The term is normally taken to refer to computer systems, or machines, which perform tasks associated with intelligent beings or which require human intelligence, the machine therefore imitating such human intelligent behavior.[8]

In its „Conversations on Intellectual Property and Artificial Intelligence",[9] WIPO has not indicated a need for a definition in its „Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence"[10] but has done so in the „Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence"[11] even in the face of a lack of consensus as to how such definition should be constructed lest what it should be, as the summary of the meeting shows.[12] WIPO finally opted for a definition that states that „„Artificial intelligence (AI)" is a discipline of computer science that is aimed at developing machines and systems that can carry out tasks considered to require human intelligence, with limited or no human intervention."[13] WIPO then proceeds to clarify that, for the purposes of its meetings, AI was referring to „narrow AI", meaning „techniques and applications programmed to perform individual tasks," with machine learning and deep learning as two subsets of AI.[14]

The European Commission, in its „Proposal for a Regulation Laying Down Harmonised Rules on Artificial

[4] *Andersen v. Stability AI Ltd*, U.S. District Court for the Northern District of California, no. 3:23-cv-00201, referenced in B. Brittain, *Lawsuits accuse AI content creators of misusing copyrighted work* in Reuters, 17.01.2023, available at *https://www.reuters.com/legal/transactional/lawsuits-accuse-ai-content-creators-misusing-copyrighted-work-2023-01-17/*, last time consulted on 06.05.2023.

[5] B.J. Copeland *artificial intelligence* in Encyclopedia Britannica, 20.03.2023, available at *https://www.britannica.com/technology/artificial-intelligence*, last time consulted on 06.05.2023.

[6] B. Marr, *The Key Definitions of Artificial Intelligence (AI) That Explain Its Importance* in Forbes, 14.02.2018, available at *https://www.forbes.com/sites/bernardmarr/2018/02/14/the-key-definitions-of-artificial-intelligence-ai-that-explain-its-importance/*, last time consulted on 06.05.2023.

[7] *Ibidem*.

[8] *Ibidem*, see also, *e.g.*, B.J. Copeland, *Artificial intelligence*, in Encyclopedia Britannica, 20.03.2023, available at *https://www.britannica.com/technology/artificial-intelligence*, last time consulted on 06.05.2023; *artificial intelligence* in Oxford Learner's Dictionaries available at *https://www.oxfordlearnersdictionaries.com/definition/english/artificial-intelligence?q=artificial+intelligence*, last time consulted on 06.05.2023; *artificial intelligence* in Merriam-Webster Dictionary, available at *https://www.merriam-webster.com/dictionary/artificial%20intelligence*, last time consulted on 06.05.2023.

[9] These have been a series of meetings organized by WIPO in its efforts to engage with its member states on questions regarding the interface between IP and AI, with the first meeting held on 27.09.2019 and another 4 sessions July and November of 2020 and a fourth one (named sixth) in September 2022.

[10] WIPO, *Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence*, 13.12.2019, available at *https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1.pdf*, last time consulted on 06.05.2023.

[11] WIPO, *Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence*, 21.05.2020, available at *https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1_rev.pdf*, last time consulted on 06.05.2023.

[12] WIPO, *Summary of Second and Third Sessions*, 04.11.2020, available at *https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_3_ge_20/wipo_ip_ai_3_ge_20_inf_5.pdf*, last time consulted on 06.05.2023, pp. 4-5.

[13] WIPO, *Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence*, 21.05.2020, available at *https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1_rev.pdf*, last time consulted on 06.05.2023, p. 3.

[14] *Ibidem*.

Intelligence (Artificial Intelligence Act)"[15] has first indicated that whatever the definition, the notion „should be clearly defined to ensure legal certainty, while providing the flexibility to accommodate future technological developments," moreover, it „should be based on the key functional characteristics of the software in particular the ability, for a given set of human-defined objectives, to generate outputs such as content, predictions, recommendations, or decisions which influence the environment with which the system interacts, be it in a physical or digital dimension."[16] Moreover, the Commission noted that „[t]he definition of AI system should be complemented by a list of specific techniques and approaches used for its development, which should be kept up-to–date in the light of market and technological developments through the adoption of delegated acts by the Commission to amend that list."[17]

The definition thus proposed, at art. 3 para. (1) is the following: „artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with." Annex I, as proposed, lists the following techniques and approaches: „(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning; (b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems; (c) Statistical approaches, Bayesian estimation, search and optimization methods."[18]

Out of the above, the AI models that were likely involved in the creation of „Heart on My Sleeve" (although the exact AI used in not known) and the ones behind Copilot/Codex and Stability Diffusion and MidJourney are all developed using machine learning. Therefore, we need to make a few remarks regarding what machine learning is and how it works.

Man-Cho So defines machine learning as „a sub-field of AI that is concerned with the automated detection of meaningful patterns in data and using the detected patterns for certain tasks"[19] which essentially involves an algorithm taking training data as input for outputting information to be used by other algorithms for prediction or decision-making.[20]

This also helps explain why the acceleration of development in the field of AI is only a more recent affair, even though research in the field has been undertaken for over 60 years. Machine learning output of a significant quality requires two cumulative factors: immense computing power (which has constantly grown up to some years back, even though in the last years no longer at the same rate[21]) and a huge pool of training data (which has required that an immense amount of data be fed into social media, e-commerce sites and the like). It is only recently that the second of the two conditions was also fulfilled to the degree to allow the compilation of training data of a sufficient quantity to allow the processed outputs to have a realistic quality so as to be significant.

However, as Man-Cho So indicates, the quality and quantity of the training data is only one of the factors that determine the success of a machine learning model. The type and formulation of the learning task and the design of the algorithm also weigh heavily on the quality of the output.[22]

The selection of the training data and the formulation of the training task are closely linked. Thus, the level of human intervention and the type of training data used depend, to a certain degree, on the type of learning that the model is envisaged for. Man-Cho So describes three such types of machine learning: supervised learning, unsupervised learning and reinforced learning.

Supervised learning briefly means that:

[15] European Commission, Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM (2021) 206, 21.04.2021, available at *https://ec.europa.eu/newsroom/dae/redirection/document/75788.pdf*, last time consulted on 06.05.2023.

[16] *Ibidem*, at preamble 6.

[17] *Ibidem*.

[18] European Commission, Annex I Artificial Intelligence Techniques and Approaches to the Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM (2021) 206, 21.04.2021, available at *https://ec.europa.eu/newsroom/dae/redirection/document/75789.pdf*, last time consulted on 06.05.2023.

[19] S. Shalev-Shwartz, S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014, referenced in A. Man-Cho So, *Technical Elements of Machine Learning for Intellectual Property Law*, p. 1, note 1, available at *https://ssrn.com/abstract=3635942*, last time consulted on 06.05.2023.

[20] A. Man-Cho So, *Technical Elements of Machine Learning for Intellectual Property Law*, pp. 1-2, available at *https://ssrn.com/abstract=3635942*, last time consulted on 06.05.2023.

[21] Moore's Law, a prediction made by Gordon Moore in 1965 that the number of transistors on a microchip would double every year while the costs are halved is generally regarded in the IT industry as a reliable indicator of the growth of computing power. The prediction has held true for a long time but the pace of progress has slowed down in recent years. The number of transistors on a microchip has been doubling every two years instead of every year since 2016.

[22] A. Man-Cho So, *Technical Elements of Machine Learning for Intellectual Property Law*, pp. 16-17, available at *https://ssrn.com/abstract=3635942*, last time consulted on 06.05.2023.

- the data on which the model is trained is labeled and the AI is trying to learn to predict the missing information in the test data based on what it was trained on;
- this, in turn, requires that all the data on which the AI is trained on be identified, selected and prepared in advance under human supervision, thus resulting in a process that becomes more and more expensive the more the quantity of training data increases (while the reliability of the model also only increases with the increase in quantity of training data, and, consequently, with cost);
- the formulation of the learning task be sufficiently precise so as to avoid over-fitting but also to allow for a discrimination in the data to be made. This also falls to the human developer of the model.

Unsupervised learning presupposes that:

- the data on which the model is trained is not labeled, but rather gathered 'as-is' from different sources, with different algorithms used for the classification of data, such as clustering;
- unsupervised learning means that the human input is, in general, limited to the selection of the type and pool of data to be gathered, thereby severely limiting costs to the process and also greatly increasing the potential amount of data to be gathered.

Reinforced learning groups together different methods of training an AI model by reference to a set of actions (defined as a policy) together with rewards and penalties for achieving a desired outcome. The model then explores variances of policies and tests them while taking stock of the outcome (reward/penalty) thereby gradually improving its course of action.

From the above, the takeaways most relevant from a copyright infringement perspective seem to be: (1) the distinction between inputs and outputs, *i.e.,* the data the model is trained on and the product of the model; and (2) the existence of different models, each with a different level of human involvement and the existence, within each such model of component algorithms, each with a different level of human input and relative importance of such. We will now proceed to consider the core issues on copyright infringement having the above takeaways in mind.

3. Issues arising in respect of copyright infringement

As indicated by WIPO, most of the issues currently identified with regards to infringement of copyright by AI refer to the inputs used for the training of the model. Thus, in the „Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence," seven out of the eight issues identified by WIPO as pertaining to „Issue 8: Infringement and Exceptions" under the heading „Copyright and Related Rights" refer to possible infringement of copyright by „the use of the data subsisting in copyright works without authorization for machine learning."[23]

Moreover, both the GitHub and Stability AI claims indicated also state an infringement by unauthorized use of copyrighted works in the training of the respective AI models (in addition to infringement by means of the output).[24]

The questions on infringement of copyright, as raised by WIPO,[25] could be summarized as follows:

- is use of the data in copyrighted works, without authorization from the right holders, for the purpose of machine learning, an infringement of the copyright in those works? If yes, should there be some balance in view of „the development of AI and the free flow of data to improve innovation in AI"?
- if the answer to the first question is in the affirmative, should there be „an explicit exception" allowing the use of such data for the training of AI applications? Should such exception be made at least for „certain acts for limited purposes, such as the use in non-commercial user-generated works or the use for research"?
- if the answer to the first question is in the affirmative, should there be a licensing scheme provided for? Should such be facilitated by means of collective management? Should remedies be limited to equitable remuneration?
- if the answer to the first question is in the affirmative, how could this infringement be detected and enforced, given the large quantity of inputs and outputs?[26] Should there be a regulatory requirement to log training data used?

As we can see from the above, the first and most important question is whether if the data used for the training of the model is contained also in works under copyright, whether such use of the data would amount to

[23] WIPO, *Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence*, 21.05.2020, pp. 8-9, available at *https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1_rev.pdf*, last time consulted on 06.05.2023.

[24] See *Doe v. GitHub Inc.*, US District Court for the Northern District of California, no. 3:22-cv-06823, para. 144-145; *Andersen v. Stability AI Ltd*, US District Court for the Northern District of California, no. 3:23-cv-00201, para. 155-157.

[25] WIPO, *Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence*, 21.05.2020, pp. 8-9, available at *https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1_rev.pdf*, last time consulted on 06.05.2023.

[26] The WIPO paper only refers to outputs but we believe that the large quantity of inputs is also just as relevant.

infringement of copyright in the underlying works.

To determine whether this would be the case, we would first have to discriminate between situations where (1) the works themselves are copyrighted or not; and (2) the data from the works is itself copyrighted or not. As Guadamuz[27] rightly points out, there are numerous types of creations which are not protected by copyright (either because they do not qualify for protection – *e.g.,* ideas, concepts, simple facts and information, etc. - or because they are in the public domain due to the expiration of the term of protection). It is evident that, where there is no copyright in the underlying works, there can be no infringement.

Therefore, where the creators of the AI limit the training data set to data from creations which are not protected by copyright, there would be no infringement by means of the inputs. However, as indicated above, there are limits and variations to the degree of human involvement in the selection of the training data set. This is most evidently the case with unsupervised learning where the models are trained on very large sets of very diverse data taken 'as such', mostly from the internet. But even with labeled sets of data used for supervised learning there is no guarantee that in such large data sets all data (however well curated) is certain to be of the non-protected kind.

In the odd case that some isn't, the fact that the user relied on the guarantee of the data set provider does not shield him from injunctive relief (or, substantially, damages to the right holder) but only, if at all assumed by the provider of the data set, provides him with a claim to be reimbursed by the latter for damage suffered as a result of the right holder's claim. Injunctive relief can be especially damaging in these cases because the model would have incorporated the 'tainted' information into its functioning thereby, should the injunction hold, the only way to respect it would be to stop the model's functioning altogether (which could also, depending on the amount of infringement, seem disproportionate).

This is made even more complicated by the fact that, once the model is accessible to the public at large, there is even less control on the inputs provided by members of the public, who could 'feed' the model with copyright protected works, which the model would then incorporate (in the specifically processed form) in its 'reference' data.

Another issue, also noted by Guadamuz,[28] is the risk that even non-protected material be accessed by means of a database thereby potentially infringing the rights in the database (either copyright or sui-generis rights).

Therefore, it would seem that relying solely on the non-protected character of the dataset used for the training of the model would not, by itself, eliminate all risk of infringement but would rather require a case-by-case analysis which would, in turn, depend on the issues of quantity of the dataset and, consequently, move the resolution of the question into the scope of the last issue mentioned above.

We thus turn to the issue of whether the use itself would fall under the scope of infringement of copyright, under the assumption that (at least some of) the material consists of works protected by copyright.

The nature of the use of the works is also fact-dependent, variations due to the type of model used being possible. As Guadamuz points out,[29] in order for the data to be analysed, a copy of the data would need to be available. Thus, the most likely right to be infringed is the right to control the reproduction of the work. This right covers not only the reproduction in whole, but also the reproduction in part, meaning – as per the CJEU's decision in Infopaq,[30] also the reproduction of just an „extract [that] contains an element of the work which, as such, expresses the author's own intellectual creation."[31] Therefore, an infringement of the right of reproduction can occur whenever protected expression is copied, without the right holder's authorization, either directly or indirectly, irrespective of the permanent or temporary nature of the copy, of the mode or form of copying, including the permanent or temporary storage of a digital copy of such.[32]

However, it would seem to us that for an infringement to be enforced against, there must be an infringer to be held accountable. Who would then be the infringer, normally identified as the one who made the copy? This will also be a fact-dependent inquiry touching upon the degree of human intervention and control in the creation of the dataset on which the model is trained. Where copies are made by a human and supplied as such for analysis to the AI, the respective human will normally be held to be the infringer. But what happens in cases where the model searches for information on the internet and builds the dataset itself? What about the

[27] A. Guadamuz, *A Scanner Darkly: Copyright Infringement in Artificial Intelligence Inputs and Outputs*, pp. 11-12, available at Electronic copy available at: *https://ssrn.com/abstract=4371204*, last time consulted on 06.05.2023.

[28] *Idem*, p. 11.

[29] *Idem*, p. 13.

[30] CJEU, Judgment of 16.07.2009, *Infopaq International A/S v. Danske Dagblades Forening* (C-5/08) in ECR-I (2009), p. 6569, ECLI:EU:C:2009:465.

[31] *Idem*, para. 48.

[32] Art. 14 of Law no. 8/1996 concerning author's right and related rights, republished in the Official Gazette of Romania no. 489/14.06.2018.

additional information the model searches for and integrates based on future prompts made by the users? Would it make a sensible difference if the human programmer had instructed the model not to copy protected works? How would the AI identify such, after all, lawyers involved in due-diligence exercises over the same works normally face significant effort hurdles to in the end achieve a rather non-unequivocal answer.

Moreover, if we were to recognize some determinative capability in the human(s) behind the AI, we would also need to take more seriously the claim for such humans to also derive copyright protection for themselves in the output of the AI.

Clearly, admitting that there could be an infringement of copyright whenever there is any reproduction of a copyrighted work in the dataset on which the AI is being trained would now have a very chilling effect on the research and development efforts surrounding AI now.

On the one hand, efforts to minimize the risk of protected expression being copied in the training of the model would force makers of such models to use only data sets which have the least possibility of containing copyrighted works. This, in turn, would severely diminish the quality of the training, and, consequently, of the model because of the (1) scarcity of information; and (2) high costs – well curated data sets will increase in price where they become the only possible source of training data and, therefore, lead to an arbitrage over the prices or, *e.g.*, over their conditions for franchising. On the other hand, increased transparency into the inner workings of the model – *e.g.,* in order to prove the limits of human control over the training of the model – makes the model much more vulnerable to attacks and, consequently, increases the use risk for everyone.

Avoiding that there is an infringement of the right of reproduction would then turn on whether a copy, as required under the law, is made. Given the wide scope of the reproduction right, as provided by law, the question would then be if the copies thus made could fit into one of the expressly provided for limitations to such a right. Guadamuz refers[33] to the 'transient copy' provisions which, in Romanian law, provide that „provisional acts of reproduction which are transient or incidental and constitute an integral and essential part of a technical process and the sole purpose of which is to enable the transmission, within a network between third parties, by an intermediary, or the lawful use of a work or other subject-matter and which are not economically significant in themselves" are outside the scope of the reproduction right (subject to the work having been made public and the use being in accordance with good practices, not affect the normal use of the work and not damage the author or right holders).[34]

Should the data set be built in such transient manner, with the data being copied solely in view of its processing and then automatically deleted, without any protected expression kept, one could deem that the copies thus made had no economic significance by themselves and therefore the exception would apply.

Moreover, art. 36^2 in the Romanian copyright law provides for a TDM exception, expressly allowing the reproduction of and extractions from protected works which are legally accessible, for the purpose of text and data mining, such reproductions and extractions being allowed to subsist for as long as necessary for the mining of the text and data. However, this exception only applies where the right holders did not expressly reserve the use of their works by appropriate means such as CMIs.

The interplay between the two provisions is not clear at this point but mention should be made that neither provides for an exception to infringement of the sui-generis right in databases, which we have mentioned above. Also, none of the exceptions deals with the use of the data for machine learning however, we believe that while the TDM exception would not shield the data mining itself from scrutiny (once the mining of text and data is complete), the inexistence of a copy where the conditions for the 'transient copy' would be deemed met, would provide a more solid defense to possible arguments of infringement by means of the information so obtained.

By reference to the issue as raised in the WIPO Conversation, we mention also that there is a wider exception expressly provided by the Romanian copyright law for the benefit of research organizations where such act for the purpose of scientific research.[35] In this case, specific indication is made that the exception also applies with regards to the sui-generis rights in databases and the copies can be retained for longer if this is justified for scientific research purposes and if adequate security measures are taken.

Thus, Romanian copyright law appears to be in a relatively good position to tackle the issues derived from the interplay between copyright and AI, as WIPO has perceived them. Future practice and case-law will however need to further clarify the scope of the provisions and provide guidance on different sets of facts.

In this context, licensing appears to be a less desirable solution, given the bottlenecks that could impede the rapid reactions needed to advance in such a dynamic field. Collective management would also be faced with

[33] A. Guadamuz, *A Scanner Darkly: Copyright Infringement in Artificial Intelligence Inputs and Outputs*, pp. 14-16, available at Electronic copy available at: *https://ssrn.com/abstract=4371204*, last time consulted on 06.05.2023.

[34] Art. 35 para. (3) of Law no. 8/1996 concerning author's right and related rights, republished, in the Official Gazette of Romania no. 489/14.06.2018.

[35] Art. 36^1 of Law no. 8/1996 concerning author's right and related rights, republished.

the high costs and risks of increased transparency in the data used for the training of the AI models and the very large quantity of data these process, thereby the cost of licensing, lest that of collective management, could vastly outweigh the benefits and, in turn, halt any development in the area.

Given the high importance the WIPO paper has given to the issue of infringement by means of the inputs, it is not surprising that, in the one issue tackling infringement by means of the outputs, the paper starts from the assumption that such infringement would require that the expression were contained in the data set with which the model was trained. Assuming the protected expression exists (and it presumably survived beyond the moment where it was needed for the purpose of the data mining operation), there will already be an infringement of the right of reproduction. A new act of reproduction could bring into play the right to make derivative works which the protected expression modified by the AI would be.

But in such a case, who is the author of the derivative work? And who is the infringer? After all, it appears that the question of authorship of AI generated works is closely related to that of infringement of copyright by AI. If we deem that someone (either the user, the programmer or the AI itself) could be liable for infringing the rights of a third party by making a derivative work based on unauthorized reproduction of protected expression, it would be natural to ask ourselves whether the derivative work so created should, in turn, be refused protection under copyright.

4. Conclusions

The paper has sought to clarify some key aspects related to possible infringement of copyright by AI and, in so doing, to raise the more salient points which could drive the analysis forward.

We have therefore tried to clarify, and give some context to, the core concept of AI and then to see how this would better delineate the perspectives from which the infringement of copyright could be further analysed. We have thus identified the variables (including the control of a human in the development and training of the model) and the two dimensions that are, in principle, relevant for the analysis of the infringement: the inputs and the outputs.

We have then proceeded to raise the issues as summarized by WIPO and to put them into a more simple framework showing how both these and the wider discussion on protectability under copyright of AI-generated works, can pan out in light of the provisions of the Romanian copyright law.

Further research is necessary to explore these paths and provide a more comprehensive assessment of the viability of arguments which exist now and to subsequently weigh in terms of both copyright policy, security policy and cultural policy the advantages and disadvantages of further molding copyright law in a way that will allow us to capture the most advantage from this very new and exciting technology.

References

- A. Man-Cho So, *Technical Elements of Machine Learning for Intellectual Property Law*, p. 1, note 1, available at *https://ssrn.com/abstract=3635942*, last time consulted on 06.05.2023;
- K. F. Milde, Jr. *Can a Computer Be an „Author" or an „Inventor"?*, in Journal of the Patent Office Society no. 51 (1969), p. 378;
- T. L. Butler, *Can a Computer Be an Author - Copyright Aspects of Artificial Intelligence*, in Hastings Communication and Entertainment Law Journal, no. 4 (1982), p. 707;
- P. Samuelson, *Allocating Ownership Rights in Computer-Generated Works*, in University of Pittsburgh Law Review no. 47 (1986), p. 1185;
- E. H. Farr, *Copyrightability of Computer-Created Works*, in Rutgers Computer and Technology Law Journal, no. 15 (1989), p. 63;
- J. Grimmelmann, *There's no Such Thing as a Computer-Authored Work*, in Columbia Journal of the Law & the Arts, no. 39 (2016);
- S. Shalev-Shwartz, S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014;
- A. Guadamuz, *A Scanner Darkly: Copyright Infringement in Artificial Intelligence Inputs and Outputs*, pp. 11-12, available at Electronic copy available at: *https://ssrn.com/abstract=4371204, last time consulted on 06.05.2023*;
- M. Sato, *Drake's AI clone is here - and Drake might not be able to stop him*, in The Verge, 01.05.2023, available at *https://www.theverge.com/2023/5/1/23703087/ai-drake-the-weeknd-music-copyright-legal-battle-right-of-publicity*, last time consulted on 06.05.2023;
- J. Coscarelli, *An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World*, in The New York Times, 19.04.2023, available at *https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html*, last time consulted on 06.05.2023;
- B.J. Copeland, *Artificial Intelligence*, in Encyclopedia Britannica, 20.03.2023, available at *https://www.britannica.com/technology/artificial-intelligence*, last time consulted on 06.05.2023;

- artificial intelligence in Oxford Learner's Dictionaries available at *https://www.oxfordlearnersdictionaries.com/definition/english/artificial-intelligence?q=artificial+intelligence*, last time consulted on 06.05.2023;
- artificial intelligence in Merriam-Webster Dictionary available at *https://www.merriam-webster.com/dictionary/artificial%20intelligence*, last time consulted on 06.05.2023;
- B. Marr, *The Key Definitions Of Artificial Intelligence (AI) That Explain Its Importance*, in Forbes, 14.02.2018, available at *https://www.forbes.com/sites/bernardmarr/2018/02/14/the-key-definitions-of-artificial-intelligence-ai-that-explain-its-importance/*, last time consulted on 06.05.2023;
- European Commission, *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* COM (2021) 206, 21.04.2021, available at *https://ec.europa.eu/newsroom/dae/redirection/document/75788.pdf*, last time consulted on 06.05.2023;
- WIPO, *Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence*, 13.12.2019, available at *https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1.pdf*, last time consulted on 06.05.2023;
- WIPO, *Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence*, 21.05.2020, available at *https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1_rev.pdf*, last time consulted on 06.05.2023;
- CJEU, Judgment of 16.07.2009, Infopaq International A/S v. Danske Dagblades Forening (C-5/08) in ECR-I (2009), p. 6569, ECLI:EU:C:2009:465S;
- Law no. 8/1996 concerning author's right and related rights, republished, in the Official Gazette of Romania no. 489/14.06.2018;
- Doe v. GitHub Inc., *US District Court for the Northern District of California*, no. 3:22-cv-06823, referenced in J. Vincent, *The lawsuit that could rewrite the rules of AI copyright*, in The Verge, 08.11.2022, available at *https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data*, last time consulted on 06.05.2023;
- Andersen *v.* Stability AI Ltd, US District Court for the Northern District of California, no. 3:23-cv-00201, referenced in B. Brittain, *Lawsuits accuse AI content creators of misusing copyrighted work*, in Reuters, 17.01.2023, available at *https://www.reuters.com/legal/transactional/lawsuits-accuse-ai-content-creators-misusing-copyrighted-work-2023-01-17/*, last time consulted on 06.05.2023.