Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

Articles

Studia Iuridica Lublinensia vol. 35, 2, 2026
DOI: 10.17951/sil.2026.35.2.219-242

Grzegorz Tylec
The John Paul II Catholic University of Lublin, Poland
ORCID: 0000-0003-2016-4523
grzegorz.tylec@kul.pl

Sebastian Kwiecień
The John Paul II Catholic University of Lublin, Poland
ORCID: 0000-0003-4858-794X
sebastian.kwiecien@kul.pl

Łukasz Sarowski
The John Paul II Catholic University of Lublin, Poland
ORCID 0000-0003-1228-9705
lukasz.sarowski@kul.pl

Adam Sadowski
University of Lodz, Poland
ORCID: 0000-0002-8608-5118
adam.sadowski@uni.lodz.pl

The Problem of Licensing Works for the Training Process of Artificial Intelligence Algorithms

Problem licencjonowania utworów na potrzeby procesu szkolenia algorytmów sztucznej inteligencji

CORRESPONDENCE ADDRESS: Grzegorz Tylec, PhD, Dr. Habil., Associate Professor, The John Paul II Catholic University of Lublin, Faculty of Social Sciences and Technology, Institute of Social Communication and Media Studies, 14 Racławickie, 20-950 Lublin, Poland; Sebastian Kwiecień, PhD, Dr. Habil., The John Paul II Catholic University of Lublin, Faculty of Law, Canon Law and Administration, Institute of Legal Sciences, 14 Racławickie, 20-950 Lublin, Poland; Łukasz Sarowski, PhD, Assistant, The John Paul II Catholic University of Lublin, Faculty of Social Sciences and Technology, Institute of Social Communication and Media Studies, 14 Racławickie, 20-950 Lublin, Poland; Adam Sadowski, PhD, Dr. Habil., Associate Professor, University of Lodz, Faculty of Management, Department of Strategy and Value-Based Management, 22/26 Matejki, 90-237 Łódź, Poland.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

ABSTRACT

One of the significant challenges in modern copyright law is determining the legal status of using copyrighted materials and related rights for training generative artificial intelligence (AI). This article examines how copyrighted works are utilized during AI training. It contrasts text and data mining (TDM) with AI training, concluding that TDM is just one of the techniques involved. Consequently, the legal frameworks in Directive 2019/790 on copyright and related rights in the Digital Single Market addressing TDM do not resolve the issue of copyright violations by generative AI developers. From a technical point of view, TDM and AI operations should be clearly distinguished from each other. Using works for AI training should be viewed as a new, unregulated field of exploitation of works and related rights beyond TDM. Therefore, in most cases, AI training will lead to the infringement of the economic rights of the creators and their successors. The authors explore the possibility of licensing works for AI training; however, the complexity in identifying which works an AI algorithm uses makes traditional copyright and related rights licensing models impractical. The solution is to introduce a statutory compensation model based on the assumption that AI producers use copyrighted works in the learning process. It is therefore postulated that a legal regulation should be introduced which would allow AI algorithm developers to use the works legally in the absence of licensing options. Therefore, the authors suggest implementing a system similar to the private copying levy found in Article 2 and Article 5(2b) of Directive 2001/29/EC. This approach is necessary to ensure fair compensation for copyright holders whose works and related rights are exploited.

Keywords: artificial intelligence; copyright law; text and data mining; data mining; private copying levy

INTRODUCTION

The literature refers to the actual process of teaching generative artificial intelligence (AI) using various terms including "learning", "teaching", and "training". This process essentially involves providing AI with information, knowledge, and experience, allowing it to simulate logical or creative thinking activities that were recently only reserved for humans.[1] Unlike the process of human learning, which has been largely regulated by law in recent years,[2] the current learning process of AI algorithms is only regulated with regard to text and data mining (TDM).[3] "Text

[1] A. Kisielewicz, *Sztuczna inteligencja i logika*, Warszawa 2017, p. 25.

[2] Even from a copyright perspective, there are clear regulations governing the use of works for teaching or scientific purposes.

[3] In EU law, Articles 3 and 4 of Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (OJ L 130/92, 17.5.2019), hereinafter: CDSM. The TDM process has also been regulated by law in some countries outside the European Union. On the regulation of TDM in Japanese law, see T. Ueno, *The Flexible Copyright Exceptions for "Non-Enjoyment" Purposes – Recent Amendment in Japan and Its Implication*, "GRUR International: Journal of European and International IP Law" 2021, vol. 70(2), pp. 145–152; A. Dermawan, *Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could*

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

mining" is a broad term that defines a range of methods and techniques for extracting useful information from document collections by identifying and utilizing patterns from unstructured text-based datasets. Formally, text mining is limited to the generation of new information based on a database of text documents that can undergo operations such as text categorization, text clustering, concept extraction, sentiment analysis, or entity relation modelling.[4] Generally, the idea behind the use of TDM is to reduce the human effort required to search through immense collections of any type of document that meets the requirements of Big Data.[5]

While IT developers will be able to create the most advanced algorithms to simulate the thinking processes of the human brain, generative AI – like the human brain – will require knowledge (factual information, experience, and education) to fully exploit its intellectual capabilities and achieve the goals set for it. This paper formulates the thesis that AI learning can be qualified as a new field of the exploitation of works, which is not yet regulated by copyright law. This assumption implies that the use of copyrighted works by AI creators for AI training may violate the law and must be regulated. With the rapid development of AI technologies, and thus of the market for such services, a prompt legal solution is required to ensure that copyright holders, whose works are used in the learning process of AI algorithms, are fairly compensated.

The need to guarantee fair remuneration to rights holders[6] for the use of their copyrighted works while considering the technical possibilities and economic interests of AI creators is indisputable. Thus, this article analyzes the lawfulness of activities aimed at teaching (training) AI algorithms and the rules for their lawful use. To solve this issue, it is crucial to answer the following question: Can the process of AI learning be limited only to the so-called TDM process? This article hypothesizes that the optimal solution would be to introduce a system for licensing the use of works by AI in the learning process. However, recognizing the significant drawbacks of this solution, which are outlined in this text, an alternative model for solving this problem is proposed. The reflections and proposals presented in this article relate to and further develop the view proposed by C. Geiger and V. Iaia in their text entitled *The Forgotten Creator: Towards a Statutory Remuneration Right*

Learn from the Japanese "Non-Enjoyment" Purposes?, "Journal of World Intellectual Property" 2023, vol. 26, pp. 1–25.

[4] M. Truyens, P. Van Eecke, *Legal Aspects of Text Mining*, "Computer Law & Security Review" 2014, vol. 30, pp. 153–170.

[5] I. Shemilt, A. Simon, G.J. Hollands, T.M. Marteau, D. Ogilvie, A. O'Mara-Eves, M.P. Kelly, J. Thomas, *Pinpointing Needles in Giant Haystacks: Use of Text Mining to Reduce Impractical Screening Workload in Extremely Large Scoping Reviews*, "Research Synthesis Methods" 2014, vol. 5(1), pp. 31–49.

[6] The term "rights holder" refers to the creator, related rights holders, and their successors in title.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

for Machine Learning of Generative AI from 2024.[7] While the present authors share the main line of these arguments, they will try to propose their own legal solutions with regard to regulating the matter at hand.

Owing to contemporary language failing to keep pace with technological developments, concepts such as AI, algorithm, or teaching/training AI algorithms remain vague and lack terminological precision in both the literature and everyday language. Moreover, due to the lack of clear criteria needed to operationalize such concepts, concepts relating to human activity, such as "learning", are not adequate for describing the process that leads to the acquisition of knowledge and experience by algorithms.

RESEARCH AND RESULTS

1. Characteristic features of the process of AI training

When approaching the solution to the problem outlined above, a few introductory remarks should be made at the outset. Artificial intelligence involves machine learning (e.g., supervised, unsupervised, deep learning), natural language processing (NLP), expert systems, image and speech processing, and planning and robotics. The problems associated with the increasing use of the above-mentioned solutions, whether in science, politics, or social debates, create the need to reflect on their potential legal implications.

The AI training process, which most commonly refers to the training of a neural network, requires copying data for subsequent transformations to achieve the model's objective. Considering that the model to be made public makes any use of a database containing copyrighted works during the training, three groups of models can be distinguished.

The first group comprises non-generative models, which are designed to recognize and assign graphical elements, such as photographs or graphics. Training these models involves using image databases to distinguish new images uploaded by users but does not result in the dissemination of images that were used for AI training.

The second group comprises generative models. An example is a language model trained on a collection of scientific articles and books. They can reproduce the source material used as the input for the neural network training process. In practice, they are trained specifically to memorize these works in their weights (a very complex representation of this data). However, because of the properties of this complex representation, it is rather rare for the model to remember an entire

[7] See C. Geiger, V. Iaia, *The Forgotten Creator: Towards a Statutory Renumeration Right for Machine Learning of Generative AI*, "Computer Law & Security Review" 2024, vol. 52(1), pp. 1–24.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

work. Instead, similar to a human memorizing a work and then recounting it from memory, the AI memorizes the meaning of the work and then paraphrases it. In theory, nothing prevents the AI from memorizing 100% of a work (just as a person with an exceptional memory can memorize an entire book and later reproduce it). When this model is used, a new work is created that, from a technological and legal point of view, can be regarded as a derivative or inspired work.

The third group of models searches for answers or compares content with material in databases consisting of copyrighted works. These models can search and load data to provide users with the required answers to their questions after processing. The AI operating in this model resembles a tool for the automatic processing and analysis of data, making it public. For example, consider a model containing a collection of texts in its database. When called, it compares the user's query to each element in this set and returns an answer regarding the number of elements it is similar to. In theory, this involves analyzing the data rather than copying them. However, as this involves briefly downloading the data into the system's operational memory to perform logical operations on them, this type of action appears to constitute an interference with copyright. Of importance in this situation is that the process has an independent economic significance.[8] It is also noteworthy that systems based on machine learning require different types of reproduction, starting with the temporal conversion of PDF files into XML, the continuous copying to create a corpus (corpora) of words or databases of training materials, and the deposition of source materials for reproducibility or verifiability purposes.[9]

Artificial intelligence pioneers such as H. Simon and D. Michie defined algorithmic learning. The former stated that "Learning reveals adaptive changes in the sense that they allow a system to perform the same task, or set of tasks, more efficiently than the first attempt".[10] The latter, in turn, noted that "The learning system utilizes external empirical data in order to create and update the foundations for the improved operation on similar data in the future and expressing these foundations in understandable and symbolic form".[11] Drawing on A. Geron, machine learning,

[8] As this process in the operation of AI has an independent economic significance, the authors opine that it cannot be regarded as falling within the exception described in Article 5(1) of Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (OJ L 167/10, 22.6.2001).

[9] See M. Senftleben, T. Margoni, D. Antal, B. Bodó, S. van Gompel, C. Handke, M. Kretschmer, J. Poort, J. Quintais, S. Schwemer, *Ensuring the Visibility and Accessibility of European Creative Content on the World Market: The Need for Copyright Data Improvement in the Light of New Technologies and the Opportunity Arising from Article 17 of the CDSM Directive*, "Journal of Intellectual Property, Information Technology and E-Commerce Law" 2022, vol. 13, pp. 67–86.

[10] See H. Simon, *Why Should Machines Learn?*, [in:] *Machine Learning: An Artificial Approach*, eds. R.S. Michalski, J.G. Carbonell, T.M. Mitchell, Palo Alto 1983, p. 25.

[11] See D. Michie, *Methodologies from Machine Learning in Data Analysis and Software*, "The Computer Journal" 1991, vol. 34, p. 45.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

which is a core area of AI research, can, in turn, be described as "the science (and art) of programming computers so they can learn from data".[12]

These definitions are the starting point for understanding the scope of applicability of the concept of learning in the field of AI. Furthermore, this concept is crucial in the context of operating the algorithms, which can be understood simply as a sequence of instructions that provide the computer with information on what it should do at any particular moment. From the perspective of the deep machine learning revolution, it has become legitimate to ask questions regarding the nature of the operations performed by the algorithms.[13] What happens during the processes between the input and output of the algorithm? How does the algorithm use the data? Why is the result of the operation the way it is?

The essence of this issue is highlighted when AI algorithms are used for projects requiring creativity and ingenuity. Examples include generative adversarial networks (GAN) technology, projects such as "The Next Rembrandt", or tools such as DALL-E image generators, Stable Diffusion, Midjourney, and ChatGPT text.[14] In the latter case, the use of machine learning-based solutions draws researchers' attention to the issue of AI-based natural language processing (NLP). ChatGPT 3.5, for example, which was based on 175 billion parameters,[15] is being developed within NLP, an interdisciplinary area of research. Briefly, NLP is an interdisciplinary area of research that combines AI with linguistics. It is concerned with the analysis, understanding, translation, and generation of natural language by computers and involves research into the use of large language models (LLMs).[16] In addition to ChatGPT, Dramatron is another example of developing this type of technology.[17]

---

[12] See A. Geron, *Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow*, Gliwice 2018, pp. 25–46.

[13] See T.J. Sejnowski, *Deep learning. Głęboka rewolucja. Kiedy sztuczna inteligencja spotyka się z ludzką*, Warszawa 2019, pp. 45–73.

[14] See L. Floridi, M. Chiriatti, *GPT-3: Its Nature, Scope, Limits, and Consequences*, "Minds and Machines" 2020, vol. 30, pp. 681–694.

[15] See D.M. Korngiebel, S.D. Mooney, *Considering the Possibilities and Pitfalls of Generative Pre-Trained Transformer 3 (GPT-3) in Healthcare Delivery*, "NPJ Digital Medicine" 2021, vol. 4, pp. 1–3.

[16] Examples of their use include the script generated for the films *Sunspring* (2016), *Solicitors* (2017), or *The First Horror Movie Written Entirely by Bots* (2021).

[17] The main page of the system using LLMs reads that "Dramatron is conceived as a writing tool and a source of inspiration and exploration for writers. To evaluate Dramatron's usability and capabilities, we engaged 15 playwrights and screenwriters in two-hour-long user study sessions to co-write scripts alongside Dramatron. (…) During the development of Dramatron and through discussions with industry professionals, we made several important observations: Dramatron is a co-writing system that has only been used in collaboration with human writers, and was not conceived or evaluated to be used autonomously; Dramatron's top-down hierarchical story generation structure does not correspond to every writer's writing process; The output of a language model may include elements of the text used to train the language model. One possible mitigation is for the human co-writer to

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

In addition to generative technologies, machine translation (MT) and computer-assisted translation (CAT) software have also been gaining popularity.[18] CATs rely on translation memory (TM) management software and language search engine (LSE) software to allow programs to reach the resources of previous translations that would coincide with, or at least be close to, the translated source text. The development of this type of technology emphasizes the importance of issues concerning the data used to train AI algorithms. The concept of data here is, on the one hand, a starting point for understanding the learning process of AI itself – its training – and, on the other hand, an area of issues not yet regulated by law. For instance, solutions such as ChatGPT and DALL-E involve problems concerning information about the data on which their algorithms were trained – whether or not they are covered by copyright – and their possible overlaps with the result of the generating software. CAT technologies can also raise concerns regarding the use of databases consisting of previous translations, which can be used by CAT and MT software without the knowledge and consent of previous translators.

2. The learning process of AI algorithms from a copyright perspective

One of the key principles of copyright law is that only the way in which a work is expressed can be protected. Therefore, discoveries, ideas, procedures, methods, principles of operation, and mathematical concepts remain unprotected[19] and can be freely used by AI without restriction. However, in order to acquire unprotected knowledge of a discovery, idea, procedure, method, operating principle, or mathematical concept, it is usually necessary to familiarize oneself with the copyrighted work in which such knowledge (information) is contained. In order to extract this knowledge, it is often necessary to enter the copyrighted material into the memory of the computer in which the AI system is operating. The reproduction (copying) of such material can thus be assumed as an element of this type of process, which may constitute copyright infringement.[20] The uncertainty as to whether or not AI

search for substrings from outputs to help to identify plagiarism; Dramatron may reproduce biases and stereotypes found in the corpus, and may generate offensive text" (GitHub, *Try Dramatron*, https://deepmind.github.io/dramatron/details.html, access: 14.7.2026).

[18] See M. Popel, M. Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, Z. Žabokrtský, *Transforming Machine Translation: A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals*, "Nature Communications" 2020, vol. 11, pp. 1–15.

[19] Berne Convention for the Protection of Literary and Artistic Works of 9 September 1886, as amended on 28 September 1979. The text is available at https://www.wipo.int/wipolex/en/text/283698 (access: 14.7.2026).

[20] See A. Škiljić, *When Art Meets Technology or Vice Versa: Key Challenges at the Crossroads of AI-Generated Artworks and Copyright Law*, "International Review of Intellectual Property and Competition Law" 2021, vol. 10, pp. 1338–1345.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

actions violate the law is one of the main reasons for the delay in the implementation of generative AI by Fortune 1,000 companies.[21]

In order to initiate the algorithm training process, AI developers are forced to perform actions that constitute an interference with the legal monopoly of rights holders, which involves downloading, recording, and digitally reproducing copyrighted material, even if this only happens in the system's operational memory and is temporary. Considering the current legal regulations, based on the machine learning problem outlined in the previous section, it is possible to pose the thesis that the use of works by AI may often be in breach of the copyright of rights holders. These breaches are concerned with both input and output data. In terms of input, the primary focus is on the activities involved in the AI training process conducted on the collection of materials, a significant proportion of which are legally protected works. The real problem in determining whether there is a breach in machine learning lies in the use of large datasets in the form of digital shadow libraries ("pirate libraries") such as Library Genesis or Libgen.[22] Libgen provides access to copyrighted works from portals such as ScienceDirect and Elsevier, which can be used to train AI models. If it is assumed that the Libgen library has a capacity of 42 terabytes of data, the scale of copyright infringement is difficult to estimate as it is for platforms such as ResearchGate, SciSpace, Sci-Hub, Science Stage, Epernicus, and Z-Library.[23]

In the scientific literature,[24] as well as in courtrooms,[25] there is currently an ongoing debate in many countries concerning the extent to which the use of legally protected works may constitute an infringement of existing copyright law. Authors from the United States were among the first in the world to seek legal protection of their rights. This conclusion is supported by the number of lawsuits that have received significant media attention, not only in the trade press. US artists seeking compensation for the unlawful use of their works have filed several lawsuits against software companies

[21] See Prifina, *Event Recap: Augmenting Consumer Experiences in the Age of Data & AI*, https://www.prifina.com/blog/event-recap-augmenting-consumer-experiences-in-the-age-of-data-ai (access: 10.7.2025).

[22] As of 4 March 2023, Libgen library contains 2.4 million non-fiction books, 80 million science magazine articles, 2.2 million fiction books, 0.4 million magazine issues, and 2 million comics strips.

[23] See H.R. Jamali, *Copyright Compliance and Infringement in ResearchGate Full-Text Journal Articles*, "Scientometrics" 2017, vol. 112(1), pp. 241–254. See also A. Maddi, D. Sapinho, *On the Culture of Open Access: The Sci-Hub Paradox*, "Scientometrics" 2023, vol. 128, pp. 5647–5658.

[24] See C. Geiger, *The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the Failed Text and Data Mining Limitations in the CDSM Directive*, [in:] *Intellectual Property and Sports: Essays in Honour of P. Bernt Hugenholtz*, eds. M. Senftleben, J. Poort, M. van Eechoud, S. Gompel, N. Helberger, Alphen aan den Rijn 2021, pp. 383–394; T. Margoni, M. Kretschmer, *A Deeper Look into the EU Text and Data Mining Exceptions: Harmonization, Data Ownership, and the Future of Technology*, "GRUR International" 2022, vol. 71, pp. 685–701; A. Dermawan, *op. cit*.

[25] Cf. court decisions on AI copyright infringement included in the fourth part of this article.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

developing AI algorithms.[26] The importance of the issue is further evidenced by the fact that representatives of the US entertainment industry participated in a hearing before the Subcommittee on Courts, Intellectual Property, and the Internet of the House of Representatives to discuss the unlawful use of works by AI and the possible imposition of an obligation on companies to compensate rights holders for their losses. Obviously, owing to the specific nature of the US legal system, the discussion has remained focused on the requirement to obtain licenses for works,[27] which is of importance for further considerations presented in this article.

The creators of AI are not only being sued by authors but also by newspaper publishers, including giants such as the "Los Angeles Times", which has already decided to take legal action against OpenAI on the grounds that ChatGPT generates texts with answers to questions based on reports and articles published in newspapers. The demands included claims for damages and a demand to destroy articles generated by ChatGPT.[28] ChatGPT is a variant of the generative pre-training transformer (GPT) model, which is trained on a massive amount of text data to generate human-like text. Similarly to other models, it does not disclose information about the source databases used to train it.[29]

Book authors have also filed a series of lawsuits claiming that texts generated by ChatGPT are, in many cases, substantially similar to their legally protected works.[30] This is because there is no doubt that a book or a scientific article is the most valuable source of knowledge and ideal input in the AI learning process, during which the AI algorithm searches the texts from a collection and extracts distinct information. The output is often already a very good simulation of natural language. This means that training material in the OpenAI and Meta datasets may originate from copyrighted works. The authors argue that, in many situations, this constitutes an infringement of copyright law.

In assessing the lawfulness of AI algorithm training activities under EU law, it is crucial to answer the question: Does the exception for TDM, introduced in the

[26] See R. Kaufman, *Some Thoughts on Five Pending AI Litigations – Avoiding Squirrels and Other AI Distractions*, 7.3.2023, https://scholarlykitchen.sspnet.org/2023/03/07/some-thoughts-on-five-pending-ai-litigations-avoiding-squirrels-and-other-ai-distractions (access: 10.7.2026).

[27] See T. Brooke, D. Garno, *Writers Continue to Resist AI Over Alleged Copyright Infringement*, 25.7.2023, https://news.bloomberglaw.com/us-law-week/writers-continue-to-resist-ai-over-alleged-copyright-infringement (access: 10.7.2026).

[28] See B. Allyn, *'New York Times' Considers Legal Action against OpenAI as Copyright Tensions Swirl*, 16.8.2023, https://www.npr.org/2023/08/16/1194202562/new-york-times-considers-legal-action-against-openai-as-copyright-tensions-swirl (access: 10.7.2026).

[29] See G. Cooper, *Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence*, "Journal of Science Education and Technology" 2023, vol. 32, pp. 444–452.

[30] See LLM litigation, *We've Filed Lawsuits Challenging ChatGPT, LLaMA, and Other Language Models for Violating the Legal Rights of Authors. Because AI Needs to Be Fair & Ethical for Everyone*, https://llmlitigation.com (access: 10.7.2026).

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

wording of Articles 3 and 4 CDSM, solve the indicated problem? These questions should be answered in the negative. This paper aims to validate the main theses that, different from TDM, the operation of AI systems in the field of input data and the related process of training algorithms should be considered new, hitherto unknown fields of exploitation of works and objects of related rights. So, one of the main theses of this paper, whose validity the authors will try to demonstrate, is that the operation of AI systems in the field of input data should be considered a new, hitherto unknown field of exploitation other than TDM. The consequence of formulating such a thesis is that the use of works to teach AI algorithms constitutes copyright infringement unless it falls within some of the exceptions provided for in the law (e.g., permitted personal use or fair use). In order to demonstrate that AI learning algorithms extend beyond TDM, a comparison of the definitions of TDM and those of AI is necessary.

According to Article 2(2) CDSM, "text and data mining" means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations. Therefore, TDM should be regarded as a process leading to the generation of information about reality based on the works undergoing this process. Essentially, any photograph or film can be treated as a source of information about reality, and TDM is a process leading to the collection of information contained in such works. According to W. Wróbel, the concept of information should be understood as a message, specific sign, sound, record, or cipher, containing a specific content.[31] Its characteristic feature is therefore that "it changes the state of consciousness from ignorance to knowledge of some part or aspect of reality".[32]

Proceeding to the definition of AI, it is first necessary to point to the AI Act,[33] which provides a normative definition in Article 3(1): "AI system means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments".

Also noteworthy is the opinion of C. Arias, who synthesized the definitions of AI found in science and developed, in the authors' opinion, one of the most

[31] W. Wróbel, *Komentarz do art. 117–277*, [in:] *Kodeks karny. Część szczególna*, ed. A. Zoll, vol. 2, Kraków 1999, pp. 968–969.

[32] J. Żyżyński, *Informacja – towar czy dobro publiczne*, [in:] *Zawód dziennikarza wobec współczesnych wyzwań*, eds. T. Kononiuk, L. Kakarenko, Warszawa 2016, p. 260.

[33] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (OJ L 2024/1689, 12.7.2024).

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

universal descriptions of the phenomenon: "AI is a field of science in which different disciplines (including computer science, computer engineering, philosophy, psychology, mathematics, neuroscience, linguistics, biology, and others) converge with the purpose of creating agents that have the ability to learn (though crudely at the moment), to create models from their learning, to make predictions based on those models and new data, and then to make autonomous decisions (or to help in human decision-making) based on those predictions and any environmental data. All of these characteristics are based on, but not copied from, our current understanding of human intelligence".[34]

A similar definition was also proposed by the European Commission's High-Level Expert Group on Artificial Intelligence[35] in its document, *A Definition of AI: Main Capabilities and Scientific Disciplines*, attempting to capture artificial intelligence from the perspective of a construct/product created by humans: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions".[36]

Based on the definitions presented above, TDM is merely an automated technique for analyzing text and digital data in order to generate information. By contrast, the definitions of AI cited above suggest that its operation extends far beyond this function and may be described as involving the generation of information that:

- is aimed at developing machines and systems;
- can carry out tasks considered to require human intelligence;
- is aimed at continuous improvement and decision-making;
- imitates actual biological cognition;
- has a purpose of generating outputs such as new content, predictions, recommendations, or decisions influencing the environments with which it interacts;
- performs analytical and physical tasks without human intervention;
- can generate (create) text, images, or other media that did not previously exist;

[34] C.R. Arias, *An Introduction to Artificial Intelligence*, 2022, https://digitalcommons.spu.edu/cgi/viewcontent.cgi?article=1169&context=works (access: 10.7.2026), p. 24.

[35] The European Commission's High-Level Expert Group on Artificial Intelligence, *A Definition of AI: Main Capabilities and Scientific Disciplines*, Brussels, 18.12.2018, https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf (access: 10.7.2026).

[36] *Ibidem*, p. 7.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

– generates new content based on or similar to the training data, but with some level of novelty (rather than just classifying or predicting the data).

To summarize, the purpose of the TDM process is only to generate information, including, i.a., patterns, trends, and correlations, whereas the functioning of AI can extend far beyond that. In particular, one of the goals of AI is to mimic human intellect, which leads to the creation of new entities (works such as texts, images, sounds). Thus, the information obtained through the TDM process is further processed by AI to form new intellectual goods.

Using the nomenclature of copyright law, AI – owing to its TDM processes – can lend an individual manner of expression to information, ideas, and operating principles extracted from the source materials. This leads to the creation of entirely new works, trademarks, or inventions. Thus, if the activity of AI is merely reduced to the creation of new information, rules, patterns, or correlations, then it may be limited to the field of exploitation, namely TDM. If, however, the data generated takes on a particular mode of expression (e.g., computer graphic, musical work, or literary text), such AI activity should fall outside the field of exploitation, that is, outside TDM.

While there is some overlap between the concepts of TDM and AI, there is, as shown above, a sphere of AI activities whose purpose must be clearly distinguished from TDM. This observation leads to the conclusion that the legal regulation concerning TDM does not solve the problem of the legality of machine learning based on copyrighted works. As aptly noted by C. Geiger and V. Iaia, the legislative solutions for TDM introduced in the 2019 CDSM were not intended to regulate generative AI, but only to address the use of one of the techniques for collecting and analyzing data used by it.[37] In the US, given the decisions in *Authors Guild v Google*[38] and *Authors Guild v HathiTrust*,[39] TDM is generally treated as conduct falling within the scope of fair use. The only debate is whether AI training can affect the copyright market and therefore such conduct falls outside the acceptable boundaries.[40]

It is worth noting here that regulations of AI Act do not create new rules within copyright law, nor can they be used to interpret the content of the DSM, as they only boil down to (a) the obligation of AI developers to comply with applicable copyright law including, i.a., Article 4(3) of Directive (EU) 2019/790 (cf. Article 53(1)(c) of the AI Act), and (b) to draw up and make publicly available a sufficiently detailed

[37] C. Geiger, V. Iaia, *op. cit.*, p. 5.

[38] U.S. Court of Appeals for the Second Circuit, *Authors Guild, Inc. v Google*, decision of 16 October 2015, No. 804 F.3d 202.

[39] U.S. Court of Appeals for the Second Circuit, *Authors Guild, Inc. v HathiTrust*, decision of 10 June 2014, No. 755 F.3d 87.

[40] Cf. Section 107 of the U.S. Copyright Act, The Act of 19 October 1976, Public Law No. 94-553, 90 Stat. 2598, effective from 1 January 1978. The text is available at https://ipmall.info/sites/default/files/hosted_resources/lipa/copyrights/Copyright1976.pdf (access: 107.2026).

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office (cf. Article 53(1)(d) of the AI Act).

The recital makes it clear that the summary should be comprehensive in its scope instead of being technically detailed, e.g. by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used. The template to be provided by the AI Office should allow the provider to provide the required summary in narrative form.

Moreover, the indicated regulations are contained in Section 2 of the AI Act "Obligations for providers of general-purpose AI models" and not in the copyright section, as there is no such section in the AI Act. A content analysis of Recitals 104–109 also leads to the conclusion that the AI Act cannot be used for any interpretation of the CDSM provisions. The wording of the explanatory memorandum shows that AI developers are only obliged to comply with applicable EU copyright law, including, i.a., respecting the decision of the right holders if they make the declaration referred to in Article 4(3) CDSM. A key argument that the provisions of the AI Act cannot be used in the discussion on the interpretation of the CDSM provisions is the statement in the last sentence of Recital 108 of the AI Act, according to which "This regulation does not affect the enforcement of copyright provisions under Union law".

Nevertheless, as noted by P. Keller, this relatively light requirement to publish high-level descriptions should be sufficient to allow third parties to determine whether model providers have trained their models on lawfully accessible data sources, as required by Article 4(1) CDSM, and therefore in compliance with EU copyright rules. According to P. Keller, the provision explicitly links the use of copyrighted works for training AI models to the TDM exception in Article 4 CDSM.[41] The direct reference to Article 4(3) CDSM in the AI Act should put an end to the discussion concerning the relationship between TDM and AI. Even if the legislator did not have these particular types of uses in mind, the legislator adopting the AI Act clearly recognizes that Article 4 CDSM applies to such uses.

In our view, however, the regulation in the AI Act described here does not resolve the issue of whether AI activity is limited to TDM only. Whether AI learning is limited solely to the TDM process or whether it constitutes a more complex phenomenon in which TDM is only one component is a question of fact and cannot be predetermined by law. The statement that AI producers are obliged to comply with Article 4(3) CDSM, in our view, does not contradict the claim that in many cases AI learning goes beyond TDM and should be treated as a new unregulated field of exploitation.

[41] See P. Keller, *A First Look at the Copyright Relevant Parts in the Final AI Act Compromise*, 11.12.2023, https://copyrightblog.kluweriplaw.com/2023/12/11/a-first-look-at-the-copyright-relevant-parts-in-the-final-ai-act-compromise (access: 10.7.2026).

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

The above observations lead to the conclusion that, under European law, even if rights holders did not avail themselves of the possibility of excluding their works[42] from the TDM process on the basis of Article 4(3) CDSM by making a declaration to that effect, the exploitation of works by AI for the purpose of teaching algorithms will still require their consent and the appropriate and equitable remuneration must be paid to them. This is because, as it is argued in the article, this is a new field of exploitation of works that is not covered by the exception. A similar conclusion was reached by C. Geiger and V. Iaia, who used the concept of the need to protect fundamental rights as a starting point.

In the context of the issue at hand, it is worth noting that the *Appeal for Action on Violations of the Berne Convention by the Application to Copying of Creative Works for AI Development of the TDM Exception in Articles 3 and 4 of the 2019 EU Directive on Copyright*, by US creators' coalition,[43] uses convincing reasoning which follows that the TDM exception adopted in Articles 3 and 4 constitutes a violation of the obligations of EU Member States as a party to the Berne Convention for the Protection of Literary and Artistic Works. This is because the exception does not meet the three-part test of Article 9(2) of the said Berne Convention. This provision allows exceptions to the exclusive right to reproduce only (1) in certain specific cases, (2) provided that such reproduction does not conflict with the normal exploitation of the work, and (3) if it does not unreasonably prejudice the author's legitimate interests.

On the legal assessment of TDM, a similar position was also expressed by the UK's Culture Media and Sport Committee, which issued a statement on 30 August 2023 that the government's original plan to exempt TDM by AI from copyright protection threatens to reduce arts and cultural production. The report argues that instead of pursuing plans for a broad TDM exemption from copyright, the government should proactively support small AI algorithm developers in acquiring licenses for the use of works.[44]

---

[42] In the literature, a view has emerged that the practical implementation of the exemption by rights holders for technical reasons will not be possible and cannot cover works published before the date of the enactment of the CDSM. See C. Geiger, G. Frosio, O. Bulayenko, *Crafting a Text and Data Mining Exception for Machine Learning and Big Data in the Digital Single Market*, [in:] *Intellectual Property and Digital Trade in the Age of Artificial Intelligence and Big Data*, eds. X. Seuba, C. Geiger, J. Penin, Geneva–Strasbourg 2018, pp. 95–112; eidem, *Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?*, "International Review of Intellectual Property and Competition Law" 2018, vol. 49, pp. 814–844.

[43] See National Writers Union, *Appeal for Action on Violations of the Berne Convention by the Application to Copying of Creative Works for AI Development of the TDM Exception in Articles 3 and 4 of the 2019 EU Directive on Copyright*, https://nwu.org/wp-content/uploads/2023/07/creators-coalition-AI-exceptions.pdf (access: 10.7.2026).

[44] See UK Parliament, *Abandon Artificial Intelligence Copyright Exemption to Protect UK Creative Industries, MPs Say*, 30.8.2023, https://committees.parliament.uk/committee/378/culture-me-

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

DISCUSSION AND CONCLUSIONS

According to general principles of copyright law, companies offering AI algorithms should be categorically required to obtain a license from rights holders even before using their works. Although the classical licensing system seems to be optimal from the point of view of rights holders' interests, the following question must be asked: To what extent is this possible in an AI environment, and if so, based on what assumptions?

When reviewing existing AI licensing proposals, three models can be distinguished: maximalist, statutory license, and statutory compensation.[45] The traditional approach, in which an entity wishing to use a work identifies a list of works and acquires the relevant rights applicable to them, should be described as the maximalist model.[46] Theoretically, in addition to this classical licensing model (referred to here as the maximalist approach), it is also possible to introduce the statutory license[47] or the statutory compensation model[48] that is preferred by the present authors. However, each solution poses a different problem. In the maximalist model, it is necessary to first identify the works that will be used before acquiring an appropriate license. However, the question that must be asked is whether it is technically possible to conclude such agreements with the obligation to identify the works that the algorithms will use in advance. The authors believe that this is not possible. Even if such agreements were to be mediated by collective management organizations, it is difficult to imagine that AI companies would be able to indicate the list of works they have used.[49] Nonetheless, attempts are being made to mandate that AI companies indicate the works they use. This is the direction of a French legislative initiative, submitted by a group of members of Parliament,

dia-and-sport-committee/news/197222/abandon-artificial-intelligence-copyright-exemption-to-protect-uk-creative-industries-mps-say (access: 10.7.2026).

[45] Cf also the solution proposed by N. Lucchi, *ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems*, "European Journal of Risk Regulation" 2023, vol. 15(3), pp. 1–23; A.M. Strowel, *ChatGPT and Generative AI Tools: Theft of Intellectual Labor?*, "International Review of Intellectual Property and Competition Law" 2023, vol. 54, pp. 491–494.

[46] This approach seems to be the guiding idea behind the French proposal described by G. Vuilletet, C. Pitollat, O. Givernet, D. Da Silva, S. Dupont, É. Chandler, E. Folest, C. Rilhac, *Proposition de loi visant à encadrer l'intelligence artificielle par le droit d'auteur*, 12.9.2023, https://www.assembleenationale.fr/dyn/16/textes/l16b1630_proposition-loi.pdf (access: 10.7.2025).

[47] C. Geiger, V. Iaia, *op. cit.*; C. Geiger, *Freedom of Artistic Creativity and Copyright Law: A Compatible Combination?*, "University of California, Irvine Law Review" 2018, vol. 8(3), p. 413.

[48] What is meant here is a legal construct modelled on the system of fair compensation for permitted personal use of so-called private copying levies. Cf. Article 2 and Article 5(2)(b) of Directive 2001/29/EC.

[49] It seems only possible to indicate databases containing the works that the AI will use, not individual works that it will use.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

which makes it compulsory for AI producers to indicate the works used by the AI algorithm in the learning process.[50]

While this undoubtedly has positive overtones in terms of respecting authors' moral rights, authors in the literature rightly point out that such a solution would be burdensome for AI creators. Some even argue that it is impossible because of the uneven threshold for the originality of works in different national jurisdictions and because only fragments of the works are used.[51] Such an approach could also limit the size of the databases of works used to train AI, which would have a negative effect on its development.[52] Even if this model can be implemented to some extent, it must still be assumed that there will be a specific sphere where source identification is not possible. Therefore, other solutions should be considered.

In the literature, C. Geiger and V. Iaia proposed the concept of a statutory license.[53] It assumes a solution that can essentially be reduced to the "permitted-but-paid" principle. With this type of solution, a provision should appear in the legal system that would, as an exception, allow certain entities, such as AI producers, to use certain types of works under the conditions indicated in this provision while requiring payment of copyright remuneration for such use. According to the present authors and contrary to the opinions expressed by C. Geiger and V. Iaia,[54] this type of legal construct should be clearly distinguished from that provided for in Article 5(2b) of Directive 2001/29/EC, which is not referred to in the Court of Justice of the European Union (CJEU) jurisprudence as a statutory license, but as a system of compensation for damages guaranteeing fair compensation to the rights holders.[55]

---

[50] G. Vuilletet, C. Pitollat, O. Givernet, D. Da Silva, S. Dupont, É. Chandler, E. Folest, C. Rilhac, *op. cit*. It is also worth adding that this project contains an original proposal for solving the copyright problem of works generated by an AI algorithm. Article 2 of the said project states that "When a work is created by artificial intelligence without direct human intervention, the only rights holders are the authors or rights holders of the works that enabled the artificial intelligence to design that work. Collective management of rights to works produced by artificial intelligence may be handled by authors' associations or other collective management organisations. These entities are entitled to represent the rights holders and receive the related remuneration for the exploitation of the copies of the works, according to the rules set out in the statutes of such companies".

[51] See C. Geiger, V. Iaia, *op. cit.*; J.B. Nordemann, J. Pukas, *Copyright Exceptions for AI Training Data – Will There Be an International Level Playing Field?*, "Journal of Intellectual Property Law Practice" 2022, vol. 17, p. 973.

[52] See K. Porayska-Pomsta, W. Holmes, S. Nemorin, *The Ethics of AI in Education*, [in:] *Handbook of Artificial Intelligence in Education*, eds. A. Mitrovic, B. du Boulay, K. Yacef, Cheltenham 2023, p. 37.

[53] See C. Geiger, V. Iaia, *op. cit.*; C. Geiger, *Freedom of Artistic Creativity…*

[54] C. Geiger, V. Iaia, *op. cit.*, p. 11.

[55] See judgment of the CJEU of 21 October 2010, C-467/08, *Padawan*, para. 40; judgment of the CJEU of 16 June 2011, C-462/09, *Stichting de Thuiskopie*, para. 24; judgment of the CJEU of 27 April 2013, C-457/11 to C-460/11, *VG Wort and Others*, para. 31; judgment of the CJEU of 11 July

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

Similar to the maximalist system, the proposal by C. Geiger and V. Iaia to introduce a statutory license construct is problematic even if AI creators would not have to inquire about the possibility to begin using the works. As opposed to the model based on private copying levies, the essence of the license model (including the so-called statutory license) is to identify all the times the work of a rights holder has been exploited in the AI learning process and to remunerate them for each use. Such an obligation to identify and indicate the extent of the works used and to calculate and remit the remuneration owed, together with a list of the works used, places a burden on AI producers. Collective management organizations would only be able to use the documents produced in this way to identify rights holders and remunerate them. Therefore, according to the authors, it will be impossible to identify each and every work used by the AI, even despite the statutory requirement, which, as has been shown, is a consequence of the very specificity of the AI training process.

Wherever the classical licensing or statutory license model falls short, the only possible solution is to introduce a statutory compensation model based on the assumption that AI producers use copyrighted works in the learning process. It is therefore postulated that a legal regulation should be introduced which would allow AI algorithm developers to use the works legally in the absence of licensing options (classical licensing model and statutory licensing). This proposed solution would draw on the exception introduced into the copyright system granting permitted use for commercial exploitation of works for AI training where there is no identifiable use of the works, or it is too costly. The introduction of such an exception would be linked to that imposed on AI producers, who would be required to pay a fee to the relevant collective management organizations representing the holders of copyright and related rights.[56]

The fee could depend on the revenue of AI producers and be reduced by any royalties, if any, paid to the rights holders by those entities.[57] This flat-rate fee, aimed at compensating the rights holders whose works were or potentially could have been used in the AI training process, would be credited to the accounts of

2013, C-521/11, *Amazon.com International Sales and Others*, para. 47; judgment of the CJEU of 10 April 2014, C-435/12, *ACI Adam and Others*, para. 50; judgment of the CJEU of 5 March 2015, C-463/12, *Copydan Båndkopi*, para. 21; judgment of the CJEU of 21 April 2016, C-572/14, *Austro-Mechana Gesellschaft zur Wahrnehmung mechanisch musikalischer Urheberrechte Gesellschaft mbH v Amazon EU Sarl and Others*, para. 19.

[56] The solution adopted presumes that if an entity using copyrighted works is not in a position to identify a list of specific works and conclude a license agreement, it may potentially benefit from a license covering all works available online. See IFPI, *European Creators and Right Holders Call for Meaningful Transparency Obligations on AI Systems to Ensure the Lawful Use of Copyright-Protected Content*, 19.7.2023, https://www.ifpi.org/european-creators-and-right-holders-call-for-meaningful-transparency-obligations-on-ai-systems-to-ensure-the-lawful-use-of-copyright-protected-content (access: 10.7.2026).

[57] Indeed, to some extent, AI producers could enter into conventional license agreements for the use of works in the AI training process in areas where this would be viable.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

the relevant collective management organizations for further distribution. The use of the collected fee would be an internal matter for the organization representing the rights holders; it could be distributed to individual rights holders according to, e.g., statistical surveys indicating which works were most frequently used by AI. It could also be earmarked for various social or support programs for the creative community if distribution to specific rights holders was not possible.

As C. Geiger and V. Iaia[58] aptly point out, the aforementioned right to fair compensation is justified by, on the one hand, the need to protect fundamental rights: the right to culture and science, freedom of expression, and freedom of artistic expression enshrined in Articles 11 and 13 EUCF,[59] Article 19 UDHR,[60] Article 27(1) UDHR, and Article 15(1)(a) and (b) ICESCR.[61] On the other hand, it is driven by the need to protect the fundamental rights of creators, as set out in Articles17(2) EUCF, Article 27(2) UDHR, and Article 15(1)(c) ICESCR. We also agree with the position of the above-mentioned authors that the solution described here can be assessed as complying with Article 9 of the Berne Convention and Article 13 TRIPS.[62] There is also a recognition that this model, which is based on the solution of Article 2 and Article 5(2b) of Directive 2001/29/EC, is not ideal. However, it would protect the financial interests of the right holders relatively quickly. At the same time, it does not inhibit the development of different licensing models, which could conceivably operate alongside each other. The full implementation of the maximalist license model or statutory license is not possible in practice. Adopting legal solutions that only seemingly safeguard the interests of rights holders and that, for technological reasons, cannot be implemented would entail the most risk. There are also no negligible costs for AI producers resulting from the need to identify works. The adoption of wrong legal solutions may thus either slow down the development of new technologies or lead to mass violations of the law.

Another tangible benefit of the proposed solution for rights holders is that it would provide them with indirect financial compensation for having their work made illegally available on the Internet by third parties without their consent. This is because AI algorithms can make use of illegal sources that are widespread on the Internet. This sphere will always be outside the licensing system. While receiving

---

[58] C. Geiger,V. Iaia, *op. cit.*, p. 17.

[59] Charter of Fundamental Rights of the European Union (OJ C 326/391, 26.10.2012).

[60] Universal Declaration of Human Rights, 10 December 1948, General Assembly Resolution 217 A. The text is available at https://digitallibrary.un.org/record/666853?v=pdf (access: 14.7.2026).

[61] International Covenant on Economic, Social and Cultural Rights, 16 December 1966, General Assembly Resolution 2200 A. The text is available at https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-economic-social-and-cultural-rights (access: 14.7.2026).

[62] Agreement on Trade-Related Aspects of Intellectual Property Rights, Annex 1C to the Marrakesh Agreement Establishing the World Trade Organization, 15 April 1994, 1869 UNTS 299. See C. Geiger, V. Iaia, *op. cit.*, p. 17.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

the compensation described here by rights holders will not legalize this kind of unlawful practice nor eliminate unlawfully distributed works from circulation, it will provide rights holders with financial compensation, if only for the use of AI in the learning process.

From a technical point of view, TDM and AI operations should be clearly distinguished from each other. Regardless of whether TDM is considered fair use, using works for AI training should be viewed as a new, unregulated field of exploitation of works and related rights beyond TDM. Therefore, in most cases, AI training will lead to the infringement of the economic rights of the creators and their successors in title.

Our position on this matter is not altered by the AI Act, which clearly indicates the obligation to apply Article 4(3) of Directive 2019/790 and to draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office. In addition, it should be noted that Articles 3 and 4 CDSM have been in force since the entry into force of the Directive. The question of making TDM in respect of works published before the entry into force of the CDSM at a time when authors could not make the relevant declarations referred to in Article 4(3) CDSM concerning their exploitation under TDM remains unresolved. There is no doubt that TDM is a separate field of exploitation of works; author contracts concluded before the date of entry into force of the CDSM do not contain provisions concerning this field of exploitation. It is therefore doubtful that rights in this field could have been acquired by publishers as regards works published before the CDSM entered into force. The doubts presented here are a further argument in favor of the need for a licensing system for works for training AI algorithms.

As we argue, even legally performing TDM under Articles 3 and 4 CDSM does not prejudge the fact that any AI algorithm learning process based on works will also be legal. When we additionally assume that, as a result of submissions made by right holders, a significant proportion of works will be excluded from the possibility of lawful TDM under Article 4(3) CDSM, the licensing of works in the AI training process seems inevitable.

Recognizing the advantages and disadvantages of the maximalist licensing model and the statutory license model in the absence of the actual option of precise identification of the works used in AI learning, this article postulates the introduction of legal regulations that will enable the statutory compensation model to take shape, and considers it advisable to allow the two solutions to operate in parallel alongside each other.

Given the subject matter described, only supranational regulations can create a transparent and fair model for compensating rights holders for their financial losses owing to the use of their works in AI training. Leaving this matter to court decisions based on the conventional provisions of copyright law would be a de facto

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

expropriation of the creators' economic rights and an action detrimental to their fundamental rights. Indeed, in ongoing litigation, significant evidentiary problems, factual imbalances between the parties, or the reliance of AI developers on the doctrine of fair use, concepts or styles that are not protected, create a real danger in terms of the lack of legal protection for the rights holders. This problem could be solved by introducing a statutory compensation model into the legal system, operating alongside classical licensing models, whereby AI producers can offset the costs incurred under conventional licensing agreements against the fee imposed on all AI producers under the statutory compensation system.

The analysis conducted above indicates that we are still at the stage of discussions, seeking legal solutions in the sphere of AI regulation, where one of the key elements is a fair balancing of the interests of rights holders and AI producers. Hopefully, this article will be an important contribution to the discussion on this issue.

REFERENCES

Literature

Cooper G., *Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence*, "Journal of Science Education and Technology" 2023, vol. 32. **DOI: https://doi.org/10.1007/s10956-023-10039-y**

Dermawan A., *Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn from the Japanese "Non-Enjoyment" Purposes?*, "Journal of World Intellectual Property" 2023, vol. 26. **DOI: https://doi.org/10.1111/jwip.12285**

Floridi L., Chiriatti M., *GPT-3: Its Nature, Scope, Limits, and Consequences*, "Minds and Machines" 2020, vol. 30. **DOI: https://doi.org/10.1007/s11023-020-09548-1**

Geiger C., *Freedom of Artistic Creativity and Copyright Law: A Compatible Combination?*, "University of California, Irvine Law Review" 2018, vol. 8(3).

Geiger C., *The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the Failed Text and Data Mining Limitations in the CDSM Directive*, [in:] *Intellectual Property and Sports: Essays in Honour of P. Bernt Hugenholtz*, eds. M. Senftleben, J. Poort, M. van Eechoud, S. Gompel, N. Helberger, Alphen aan den Rijn 2021.

Geiger C., Frosio G., Bulayenko O., *Crafting a Text and Data Mining Exception for Machine Learning and Big Data in the Digital Single Market*, [in:] *Intellectual Property and Digital Trade in the Age of Artificial Intelligence and Big Data*, eds. X. Seuba, C. Geiger, J. Penin, Geneva–Strasbourg 2018.

Geiger C., Frosio G., Bulayenko O., *Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?*, "International Review of Intellectual Property and Competition Law" 2018, vol. 49. **DOI: https://doi.org/10.1007/s40319-018-0722-2**

Geiger C., Iaia V., *The Forgotten Creator: Towards a Statutory Renumeration Right for Machine Learning of Generative AI*, "Computer Law & Security Review" 2024, vol. 52(1). **DOI: https://doi.org/10.1016/j.clsr.2023.105925**

Geron A., *Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow*, Gliwice 2018.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

Jamali H.R., *Copyright Compliance and Infringement in ResearchGate Full-Text Journal Articles*, "Scientometrics" 2017, vol. 112(1). **DOI: https://doi.org/10.1007/s11192-017-2291-4**

Kisielewicz A., *Sztuczna inteligencja i logika*, Warszawa 2017.

Korngiebel D.M., Mooney S.D., *Considering the Possibilities and Pitfalls of Generative Pre-Trained Transformer 3 (GPT-3) in Healthcare Delivery*, "NPJ Digital Medicine" 2021, vol. 4.

Lucchi N., *ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems*, "European Journal of Risk Regulation" 2023, vol. 15(3). **DOI: https://doi.org/10.1017/err.2023.59**

Maddi A., Sapinho D., *On the Culture of Open Access: The Sci-Hub Paradox*, "Scientometrics" 2023, vol. 128. **DOI: https://doi.org/10.1007/s11192-023-04792-5**

Margoni T., Kretschmer M., *A Deeper Look into the EU Text and Data Mining Exceptions: Harmonization, Data Ownership, and the Future of Technology*, "GRUR International" 2022, vol. 71.

Michie D., *Methodologies from Machine Learning in Data Analysis and Software*, "The Computer Journal" 1991, vol. 34.

Nordemann J.B., Pukas J., *Copyright Exceptions for AI Training Data – Will There Be an International Level Playing Field?*, "Journal of Intellectual Property Law Practice" 2022, vol. 17.

Popel M., Tomkova M., Tomek J., Kaiser Ł., Uszkoreit J., Bojar O., Žabokrtský Z., *Transforming Machine Translation: A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals*, "Nature Communications" 2020, vol. 11.

Porayska-Pomsta K., Holmes W., Nemorin S., *The Ethics of AI in Education*, [in:] *Handbook of Artificial Intelligence in Education*, eds. A. Mitrovic, B. du Boulay, K. Yacef, Cheltenham 2023.

Sejnowski T.J., *Deep learning. Głęboka rewolucja. Kiedy sztuczna inteligencja spotyka się z ludzką*, Warszawa 2019.

Senftleben M., Margoni T., Antal D., Bodó B., Gompel S. van, Handke C., Kretschmer M., Poort J., Quintais J., Schwemer S., *Ensuring the Visibility and Accessibility of European Creative Content on the World Market: The Need for Copyright Data Improvement in the Light of New Technologies and the Opportunity Arising from Article 17 of the CDSM Directive*, "Journal of Intellectual Property, Information Technology and E-Commerce Law" 2022, vol. 13.

Shemilt I., Simon A., Hollands G.J., Marteau T.M., Ogilvie D., O'Mara-Eves A., Kelly M.P., Thomas J., *Pinpointing Needles in Giant Haystacks: Use of Text Mining to Reduce Impractical Screening Workload in Extremely Large Scoping Reviews*, "Research Synthesis Methods" 2014, vol. 5(1). **DOI: https://doi.org/10.1002/jrsm.1093**

Simon H., *Why Should Machines Learn?*, [in:] *Machine Learning: An Artificial Approach*, eds. R.S. Michalski, J.G. Carbonell, T.M. Mitchell, Palo Alto 1983.

Strowel A.M., *ChatGPT and Generative AI Tools: Theft of Intellectual Labor?*, "International Review of Intellectual Property and Competition Law" 2023, vol. 54. **DOI: https://doi.org/10.1007/s40319-023-01321-y**

Škiljić A., *When Art Meets Technology or Vice Versa: Key Challenges at the Crossroads of AI-Generated Artworks and Copyright Law*, "International Review of Intellectual Property and Competition Law" 2021, vol. 10. **DOI: https://doi.org/10.1007/s40319-021-01119-w**

Truyens M., Van Eecke P., *Legal Aspects of Text Mining*, "Computer Law & Security Review" 2014, vol. 30. **DOI: https://doi.org/10.1016/j.clsr.2014.01.009**

Ueno T., *The Flexible Copyright Exceptions for "Non-Enjoyment" Purposes – Recent Amendment in Japan and Its Implication*, "GRUR International: Journal of European and International IP Law" 2021, vol. 70(2).

Wróbel W., *Komentarz do art. 117–277*, [in:] *Kodeks karny. Część szczególna*, ed. A. Zoll, vol. 2, Kraków 1999.

Żyżyński J., *Informacja – towar czy dobro publiczne*, [in:] *Zawód dziennikarza wobec współczesnych wyzwań*, eds. T. Kononiuk, L. Kakarenko, Warszawa 2016.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

Online sources

Allyn B., *'New York Times' Considers Legal Action against OpenAI as Copyright Tensions Swirl*, 16.8.2023, https://www.npr.org/2023/08/16/1194202562/new-york-times-considers-legal-action-against-openai-as-copyright-tensions-swirl (access: 10.7.2026).

Arias C.R., *An Introduction to Artificial Intelligence*, 2022, https://digitalcommons.spu.edu/cgi/viewcontent.cgi?article=1169&context=works (access: 10.7.2026).

Brooke T., Garno D., *Writers Continue to Resist AI Over Alleged Copyright Infringement*, 25.7.2023, https://news.bloomberglaw.com/us-law-week/writers-continue-to-resist-ai-over-alleged-copyright-infringement (access: 10.7.2026).

GitHub, *Try Dramatron*, https://deepmind.github.io/dramatron/details.html (access: 14.7.2026).

IFPI, *European Creators and Right Holders Call for Meaningful Transparency Obligations on AI Systems to Ensure the Lawful Use of Copyright-Protected Content*, 19.7.2023, https://www.ifpi.org/european-creators-and-right-holders-call-for-meaningful-transparency-obligations-on-ai-systems-to-ensure-the-lawful-use-of-copyright-protected-content (access: 10.7.2026).

Kaufman R., *Some Thoughts on Five Pending AI Litigations – Avoiding Squirrels and Other AI Distractions*, 7.3.2023, https://scholarlykitchen.sspnet.org/2023/03/07/some-thoughts-on-five-pending-ai-litigations-avoiding-squirrels-and-other-ai-distractions (access: 10.7.2026).

Keller P., *A First Look at the Copyright Relevant Parts in the Final AI Act Compromise*, 11.12.2023, https://copyrightblog.kluweriplaw.com/2023/12/11/a-first-look-at-the-copyright-relevant-parts-in-the-final-ai-act-compromise (access: 10.7.2026).

LLM litigation, *We've Filed Lawsuits Challenging ChatGPT, LLaMA, and Other Language Models for Violating the Legal Rights of Authors. Because AI Needs to Be Fair & Ethical for Everyone*, https://llmlitigation.com (access: 10.7.2026).

National Writers Union, *Appeal for Action on Violations of the Berne Convention by the Application to Copying of Creative Works for AI Development of the TDM Exception in Articles 3 and 4 of the 2019 EU Directive on Copyright*, https://nwu.org/wp-content/uploads/2023/07/creators-coalition-AI-exceptions.pdf (access: 10.7.2026).

Prifina, *Event Recap: Augmenting Consumer Experiences in the Age of Data & AI*, https://www.prifina.com/blog/event-recap-augmenting-consumer-experiences-in-the-age-of-data-ai (access: 10.7.2025).

The European Commission's High-Level Expert Group on Artificial Intelligence, *A Definition of AI: Main Capabilities and Scientific Disciplines*, Brussels, 18.12.2018, https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf (access: 10.7.2026).

UK Parliament, *Abandon Artificial Intelligence Copyright Exemption to Protect UK Creative Industries, MPs Say*, 30.8.2023, https://committees.parliament.uk/committee/378/culture-media-and-sport-committee/news/197222/abandon-artificial-intelligence-copyright-exemption-to-protect-uk-creative-industries-mps-say (access: 10.7.2026).

Vuilletet G., Pitollat C., Givernet O., Da Silva D., Dupont S., Chandler É., Folest E., Rilhac C., *Proposition de loi visant à encadrer l'intelligence artificielle par le droit d'auteur*, 12.9.2023, https://www.assembleenationale.fr/dyn/16/textes/l16b1630_proposition-loi.pdf (access: 10.7.2025).

Legal acts

Agreement on Trade-Related Aspects of Intellectual Property Rights, Annex 1C to the Marrakesh Agreement Establishing the World Trade Organization, 15 April 1994, 1869 UNTS 299.

Berne Convention for the Protection of Literary and Artistic Works of 9 September 1886, as amended on 28 September 1979.

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

Charter of Fundamental Rights of the European Union (OJ C 326/391, 26.10.2012).
Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (OJ L 167/10, 22.6.2001).
Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (OJ L 130/92, 17.5.2019).
International Covenant on Economic, Social and Cultural Rights, 16 December 1966, General Assembly Resolution 2200 A.
Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (OJ L 2024/1689, 12.7.2024).
U.S. Copyright Act, The Act of 19 October 1976, Public Law No. 94-553, 90 Stat. 2598, effective from 1 January 1978.
Universal Declaration of Human Rights, 10 December 1948, General Assembly Resolution 217 A.

Case law

Judgment of the CJEU of 21 October 2010, C-467/08, *Padawan*, para. 40.
Judgment of the CJEU of 16 June 2011, C-462/09, *Stichting de Thuiskopie*, para. 24.
Judgment of the CJEU of 27 April 2013, C-457/11 to C-460/11, *VG Wort and Others*, para. 31.
Judgment of the CJEU of 11 July 2013, C-521/11, *Amazon.com International Sales and Others*, para. 47.
Judgment of the CJEU of 10 April 2014, C-435/12, *ACI Adam and Others*, para. 50.
Judgment of the CJEU of 5 March 2015, C-463/12, *Copydan Båndkopi*, para. 21.
Judgment of the CJEU of 21 April 2016, C-572/14, *Austro-Mechana Gesellschaft zur Wahrnehmung mechanisch musikalischer Urheberrechte Gesellschaft mbH v Amazon EU Sarl and Others*, para. 19.
U.S. Court of Appeals for the Second Circuit, *Authors Guild, Inc. v HathiTrust*, decision of 10 June 2014, No. 755 F.3d 87.
U.S. Court of Appeals for the Second Circuit, *Authors Guild, Inc. v Google*, decision of 16 October 2015, No. 804 F.3d 202.

ABSTRAKT

Jednym z istotnych wyzwań współczesnego prawa autorskiego jest odpowiedź na pytanie, czy wykorzystywanie materiałów chronionych prawem autorskim i prawami pokrewnymi do szkolenia generatywnej sztucznej inteligencji (AI – *artificial intelligence*) jest dozwolone. Artykuł zawiera analizę procesu wykorzystywania chronionych prawem autorskim utworów do szkolenia algorytmów AI. Dokonano porównania procesu *text and data mining* (TDM) z procesem szkolenia algorytmów AI. Na podstawie uzyskanych wyników stwierdzono, że TDM to tylko jedna z wielu technik zaangażowanych w ten proces. Następnie ustalono, że ramy prawne określone w dyrektywie 2019/790 o prawach autorskich i prawach pokrewnych na jednolitym rynku cyfrowym dotyczące TDM nie rozwiązują kwestii naruszeń praw autorskich przez deweloperów generatywnej AI. Z technicznego punktu widzenia zjawiska, którymi są TDM oraz szkolenie generatywnej AI powinny być wyraźnie

Pobrane z czasopisma Studia Iuridica Lublinensia **http://studiaiuridica.umcs.pl**
Data: 21/11/2026 12:12:36

rozróżniane. Wykorzystanie utworów do szkolenia AI powinno być postrzegane jako nowe, nieuregulowane dotychczas w prawie autorskim, pole eksploatacji utworów i praw pokrewnych, inne niż TDM. W oparciu o analizę możliwości licencjonowania utworów do szkolenia AI autorzy doszli do wniosku, że w wielu sytuacjach nie jest możliwa identyfikacja tego, które utwory są wykorzystywane przez algorytmy AI, co uniemożliwia zastosowanie tradycyjnych modeli licencjonowania praw autorskich i praw pokrewnych. Jako rozwiązanie tej sytuacji zaproponowano wprowadzenie ustawowego modelu kompensacyjnego opartego na założeniu, że producenci AI w sposób bezumowny wykorzystują chronione prawem autorskim utwory w procesie uczenia algorytmów. Postulaty obejmują wprowadzenie regulacji prawnej, która umożliwiłaby twórcom algorytmów AI legalne wykorzystanie utworów i przedmiotów praw pokrewnych, gdy nie ma technicznej możliwości ich licencjonowania. W związku z tym autorzy proponują wprowadzenie systemu podobnego do opłaty reprograficznej, o której mowa w art. 2 i art. 5 ust. 2b dyrektywy 2001/29/EC. Takie podejście jest konieczne, aby zapewnić uczciwe wynagrodzenie dla posiadaczy praw autorskich, których utwory i prawa pokrewne są wykorzystywane.

Słowa kluczowe: sztuczna inteligencja; prawo autorskie; eksploracja tekstu i danych; eksploracja danych; opłata reprograficzna

UMCS