https://doi.org/10.22364/jull.17.03

Natural Language, Legal Hurdles: Navigating the Complexities in Natural Language Processing Development and Application

Mg. iur. **Ilya Ilin**
Faculty of Law, University of Tartu
Doctoral student
E-mail: *ilya.ilin@ut.ee*

PhD **Aleksei Kelli**
Faculty of Law, University of Tartu
Professor of Intellectual Property Law
E-mail: *aleksei.kelli@ut.ee*

This article delves into the legal challenges faced in developing and deploying Natural Language Processing (NLP) technologies, focusing particularly on the European Union's legal framework, especially the DSM Directive, the InfoSoc Directive, and the Artificial Intelligence Act. It addresses the legal status and accessibility of language data and the development of NLP technologies under both contractual and exception-based models. The authors acknowledge the partial truth in the saying, "US innovates, China replicates, and the EU regulates". Although Europe's AI sector is a global competitor and its strict regulations ensure ethical standards and data protection, these regulations might not necessarily boost competitiveness. Such stringent regulations can introduce complexities that may inhibit innovation relative to regions with more lenient policies.

Keywords: natural language processing, copyright, artificial intelligence.

Contents

Introduction

Natural language processing (NLP) represents a specialized branch of artificial intelligence (AI) focused on enabling machines to interpret and interact with human language. This technology empowers computers to comprehend, recognize, process, and produce spoken and written forms of human communication.[1] Article 3(1) of the Artificial Intelligence Act[2] defines an AI system as

> *a machine-based system designed to operate with varying levels of autonomy, 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.*

The development of artificial intelligence (AI), particularly in the realm of NLP, has sparked numerous legal debates globally. This article seeks to add to these scholarly conversations, presenting the authors' initial perspective on the prevailing legal issues, especially within the ambit of intellectual property[3] (IP) law. Emphasis is placed predominantly on issues pertaining to copyright and related (neighbouring) rights.[4] The primary emphasis of this manuscript is on copyright, while due to the focus of the paper and space limitations, the systematic analysis of issues pertaining to trade secrets and personal data is absent. To a certain degree, the paper engages with the contractual dimensions of NLP. Specifically, it scrutinizes how the contractual terms governing NLP applications delineate the parameters for the utilization of user-generated input data, such as prompts and descriptions. Moreover, the relevance of contractual terms extends to the outputs of NLP, as they may confer *de facto* ownership rights and impose restrictions on use.

This paper expands its analysis beyond the mere development of NLP applications to explore the intricacies of their usage. It focuses on two main areas: firstly, the input aspects of NLP, which involve accessing and using language data[5] protected by copyright and related rights; and secondly, the legal status of the outcomes generated by NLP applications.

The authors highlight the lack of clarity of the existing legal framework in the EU to effectively manage the distinctive challenges brought forth by the evolution and

---

[1] *Barthélemy, F., Ghesquière, N., Loozen, N. et al.* Natural language processing for public services. European Commission, Directorate-General for Digital Services. Publications Office of the European Union, 2022. Available: https://data.europa.eu/doi/10.2799/304724 [last viewed 14.04.2024].

[2] European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html [last viewed 14.04.2024].

[3] Intellectual property (IP) includes rights resulting from intellectual activity in the industrial, scientific, literary or artistic fields. Article 2 (viii) of the Convention Establishing the World Intellectual Property Organization (as amended on 28 September 1979). Available: https://www.wipo.int/wipolex/en/text/283854 [last viewed 14.04.2024]. IP is traditionally divided into three main categories: copyright, related rights and industrial property.

[4] In the context of the article, the term "copyright" is frequently used to encompass related rights.

[5] According to Article 2 (1) of the Data Governance Act, "'data' means any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audiovisual recording". Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). OJ L 152, 3.6.2022, pp. 1–44. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868&qid=1712322025563 [last viewed 14.04.2024].

implementation of NLP applications. This inadequacy is particularly evident in the legal ambiguity surrounding the training data (input) and the generated content (output). Such a lack of legal clarity could potentially have an adverse effect on the development and utilization of NLP technologies.

The authors argue that the regulatory framework for NLP, as outlined by the DSM Directive[6] and the InfoSoc Directive[7] in conjunction with the Artificial Intelligence Act, supports NLP development under a contractual model that imposes a remuneration obligation on AI service providers. This stifles innovation within Europe. Detailed arguments are presented below.

Before the adoption of the DSM Directive, NLP development primarily relied on the exceptions for temporary acts of reproduction, personal use, quotation rights, and research as defined in the InfoSoc Directive. The authors maintain that these exceptions remain pertinent today.

The DSM Directive introduced two exceptions for Text and Data Mining (TDM): the general-purpose TDM exception (Article 4) and the TDM exception for research (Article 3). However, both of these exceptions come with limitations that restrict the development and use of NLP technologies in Europe.

A common legal challenge for both TDM exceptions involves the legal uncertainty surrounding the concept of lawful access. It would benefit the development of NLP, if the concept of lawful access did not necessarily include access to a legal source. However, the authors argue that such an interpretation would conflict with the three-step test.

The main challenge of the general-purpose TDM (Text and Data Mining) exception concerns the rightsholder's opt-out right, which is reinforced by the transparency obligation of the AI Act. The TDM exception for research does not involve the opt-out right, but the primary issue is whether a provider of AI services can use a language model developed under this exception outside research settings. The authors express doubts about this possibility.

Legal challenges related to the output of NLP include the legal status of prompts, and further issues concerning the output of NLP encompass ethical and ownership implications. From an ethical perspective, the use of AI is not inherently negative; however, a transparency obligation should be enforced. Due to the absence of property rights covering NLP output, the contractual standard terms dictated by AI service providers prevail.

This research is mainly confined to the European Union (EU) legal framework. Given the United States' status as a leading knowledge-based economy and the international consensus on certain elements of its copyright system manifested among others in the TRIPS Agreement[8], examples from the US are also incorporated. The cases chosen to highlight legal issues extend beyond spoken and written language,

---

[6] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (DSM Directive). OJ L 130, 17.5.2019, pp. 92–125. Available: http://data.europa.eu/eli/dir/2019/790/oj [last viewed 14.04.2024].

[7] 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 (InfoSoc Directive) OJ L 167, 22.6.2001, pp. 10–19. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32001L0029&qid=1712245927823 [last viewed 14.04.2024].

[8] Agreement on Trade-Related Aspects of Intellectual Property Rights, amended on 23 January 2017 (TRIPS Agreement). Available: https://www.wto.org/english/docs_e/legal_e/31bis_trips_01_e.htm [last viewed 14.04.2024].

encompassing audio, video, and image materials. The analysis builds upon the authors' preceding work rather than commencing from scratch.[9]

This article's sentence structure and communicative effectiveness were enhanced with the assistance of *OpenAI's ChatGPT* and *Grammarly*. However, it is important to clarify that the substantial analytical content and core insights were exclusively the work of the authors.

1. Natural Language Processing (NLP): The technological perspective

NLP lacks a universal legal definition, prompting discussions on its relationship with the broader AI concept and its role within AI legal frameworks. A concise overview of NLP technology will be provided to clarify this linkage.

NLP is not a new technology; its roots trace back to the 1950s, when researchers began exploring ways to enable machines to understand and interact with human language.[10] Over the decades, NLP has undergone significant development, with milestones such as the advent of rule-based systems in the 1960s and the rise of statistical models in the 1990s.[11] In recent years, the prominence of neural networks, particularly deep learning models, has marked a substantial breakthrough in NLP, showcasing the capability to handle extensive text data, comprehend intricate language patterns, and perform various tasks like machine translation, sentiment analysis, and text classification.[12] Concurrently, the emerging field of "generative AI" has gained significance. Large Language Models (LLMs) models like *GPT-3* exemplify the progress in this domain by generating human-like text and contributing to significant advancements in natural language understanding and content generation.[13]

In general, the development of NLP technology involves several stages, starting with language data collection, where vast amounts of text data are gathered from diverse sources. This is followed by pre-processing to clean and prepare the data (data preprocessing).[14] Next, the model training stage involves using machine learning

[9] See, e.g., *Kelli, A., Tavast, A., Lindén, K.* Building a Chatbot: Challenges under Copyright and Data Protection Law. In: Contracting and Contract Law in the Age of Artificial Intelligence. *Ebers, M., Poncibò, C., Zou, M.* (eds). Bloomsbury Publishing, 2022, pp. 115–134; *Ilin, I.* Legal Regime of the Language Resources in the Context of the European Language Technology Development. In: Human Language Technology. Challenges for Computer Science and Linguistics. Springer Nature Switzerland AG: Lecture Notes in Computer Science, 13212, 2022, pp. 367–376. Available: https://doi.org/10.1007/978-3-031-05328-3_24 [last viewed 14.04.2024].

[10] *King, M. R., & ChatGPT.* A conversation on artificial intelligence, chatbots, and plagiarism in higher education. Cellular and molecular bioengineering, 16(1), 2023, pp. 1–2. Available: https://doi.org/10.1007/s12195-022-00754-8 [last viewed 14.04.2024]; *Weizenbaum, J.* ELIZA – a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 1966, pp. 36–45. Available: https://doi.org/10.1145/365153.365168 [last viewed 14.04.2024].

[11] *Feng, Z.* Past and Present of Natural Language Processing. In: Formal Analysis for Natural Language Processing: A Handbook. Singapore: Springer, 2023, pp. 3–48. Available: https://doi.org/10.1007/978-981-16-5172-4_1 [last viewed 14.04.2024].

[12] *Imamguluyev, R.* The Rise of GPT-3: Implications for Natural Language Processing and Beyond. International Journal of Research Publication and Reviews, 4(3), 2023, pp. 4893–4903. Available: https://doi.org/10.55248/gengpi.2023.4.33987 [last viewed 14.04.2024].

[13] *Foster, D.* Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play (Japanese Version). O'Reilly Media Incorporated, 2019, pp. 139–140.

[14] *Goldberg, Y.* Features for Textual Data. In: Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies. Springer, Cham. 2017, pp. 65–76. Available: https://doi.org/10.1007/978-3-031-02165-7_6 [last viewed 14.04.2024].

algorithms to learn patterns and structures in the data (model training).[15] After training, the model is evaluated and fine-tuned to improve its performance on language tasks (evaluation and fine-tuning). Finally, deployment involves making the model available for use in applications, with ongoing monitoring and updates based on feedback and advancements in technology (deployment and continuous improvement).

The creation of language datasets inherently involves the extensive use of language data. These data include textual data (e.g. written texts, transcribed speech, annotated sentences), speech data (e.g. audio recordings, phonetic and prosodic annotations), and multimodal data (e.g. image and text pairs, video and text, audio-text alignments).[16]

Following the development and deployment of a language model, the capacity to generate textual outputs accrues. These outputs may derive from discernible patterns and insights acquired during the training phase or be guided by specific instructions, commonly called prompts. For example, language models like *ChatGPT* generate textual outputs by interpreting user prompts, which guide the model in addressing specific tasks through semantic analysis and contextual understanding.[17] These prompts, serving as user inputs, lead to the generation of outputs through NLP algorithms, primarily transformer-based models[18], enabling the model to produce coherent and contextually appropriate responses. Put differently, prompts are converted into the resulting output. From a legal standpoint, the status of prompts in the realm of AI-generated content entails various complexities, predominantly concerning IP rights, contractual duties, and regulatory compliance.

At the same time, AI-generated content can be delivered across a spectrum of formats. These include text-based outputs like articles, stories, chatbot responses, and code snippets. Additionally, NLP models can generate structured data, such as tables or datasets, and contribute to interactive experiences like virtual assistants or personalized recommendations. The format of the generated content depends on factors such as input data, the capabilities of the NLP model, and the intended use case or platform.

2. Legal challenges related to the development of NLP

2.1. Language data access: copyright and related rights protection

Language data, crucial for NLP development, comes from a wide range of sources. This includes social media, speech and audio content websites, online publications, and sharing platforms like *GitHub*, as well as specialized repositories like *CLARIN*[19]

---

[15] *Zhou, M., Duan, N., Liu, S., & Shum, H. Y.* Progress in neural NLP: modeling, learning, and reasoning. Engineering, 6(3), 2020, pp. 275–290. Available: https://doi.org/10.1016/j.eng.2019.12.014 [last viewed 14.04.2024].

[16] *Dash, N. S., & Arulmozi, S.* History, features, and typology of language corpora. Singapore: Springer, 2018, p. 291. Available: https://doi.org/10.1007/978-981-10-7458-5 [last viewed 14.04.2024].

[17] *Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H.* Reframing Instructional Prompts to GPTk's Language. 2021. Available: https://doi.org/10.48550/arXiv.2109.07830 [last viewed 14.04.2024].

[18] *Mayer, C. W., Ludwig, S., & Brandt, S.* Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models. Journal of Research on Technology in Education, 55(1), 2023, pp. 125–141. Available: https://doi.org/10.1080/15391523.2022.2142872 [last viewed 14.04.2024].

[19] Common Language Resources and Technology Infrastructure (CLARIN). Available: https://www.clarin.eu/content/clarin-nutshell [last viewed 14.04.2024].

and OpenCorpora, governmental[20] and institutional databases. Each source presents distinct opportunities and challenges regarding accessibility, quality, and legal considerations, highlighting the complex process of acquiring language data for NLP projects. The collection and processing of this data give rise to numerous legal considerations, covering areas such as intellectual property rights, data protection regulations, contract law, tort law, and other relevant legal fields.

Web scraping, a common data collection method, carries legal risks, including contract violations and copyright infringement. Illegally scraping copyrighted content or violating website terms of service may result in legal consequences, emphasizing the importance of careful data acquisition practices.[21] The issue is highlighted in the case of Doe et al. v. *GitHub*, *Inc.* et al.,[22] where *Microsoft*, *GitHub*, and *OpenAI* are accused of extensive "software piracy" through their AI coding assistant, *GitHub Copilot*. *GitHub Copilot* learns from publicly available code repositories scraped from the internet. Plaintiffs argue that by using this data, the defendants violated the rights of creators who shared code under open-source licenses, such as MIT and GPL, which require crediting the author. Furthermore, the defendants are accused of violating *GitHub's* terms of service and privacy policies.

Data sharing and reuse also bring up data ownership, integrity rights, and ethics issues. Recent discourse has highlighted data poisoning, as seen in the Nightshade article,[23] where artists counteract generative AI's unwanted content harvesting by intentionally manipulating data. This emphasizes the critical need to tackle legal issues related to copyrights and related rights legislation, such as whether data poisoning to prevent its harvesting could constitute a technological protection measure within the meaning of the InfoSoc Directive.

From the perspective of IP law, language data as such is often eligible for protection under copyright and related rights legislation. According to Article 2(1) of the Berne Convention[24], "'literary and artistic works' shall include every production in the literary, scientific and artistic domain". The Court of Justice of the European Union (CJEU), citing several cases, reiterates:

> *The concept of 'work' that is the subject of all those provisions constitutes, as is clear from the Court's settled case-law, an autonomous concept of EU law which must be interpreted and applied uniformly, requiring two cumulative conditions to be satisfied. First, that concept entails that there exists an original subject matter, in the sense of being the author's own intellectual*

[20] Data Governance Act plays an essential role in the re-use of data held by public sector bodies and data intermediation services. For further discussion, see *Kamocki, P., Linden, K., Puksas, A., Kelli, A.* EU Data Governance Act: Outlining a Potential Role for CLARIN. In: CLARIN Annual Conference 2022, *Erjavec, T., Eskevich, M.* (eds). Linköping Electronic Conference Proceedings, 2023, pp. 57–65. Available: https://doi.org/10.3384/ecp198006 [last viewed 14.04.2024].

[21] *Pagallo, U., & Sciolla, J. C.* Anatomy of web data scraping: ethics, standards, and the troubles of the law. European Journal of Privacy Law & Technologies, No. 2, 2023, pp. 6–7, 9–10. Available: https://dx.doi.org/10.2139/ssrn.4707651[last viewed 14.04.2024].

[22] Class Action Complaint and Demand for Jury Trial in case No. 3:22-cv-06823, Doe et al. v. GitHub, Inc. et al., United States District Court for the Northern District of California. Available: https://githubcopilotlitigation.com/pdf/06823/1-0-github_complaint.pdf [last viewed 14.04.2024].

[23] *Heikkila, M.* This new data poisoning tool lets artists fight back against generative AI. 2023. Available: https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai [last viewed 14.04.2024].

[24] Berne Convention for the Protection of Literary and Artistic Works, signed at Berne on 9 September 1886 (Berne Convention). Available: https://www.wipo.int/wipolex/en/text/283698 [last viewed 14.04.2024].

> *creation. Second, classification as a work is reserved to the elements that are the expression of such creation.*[25]

CJEU also holds that the significant labour and skill cannot, as such, justify copyright protection, if they do not express any originality.[26]

Since copyright-protected work is an autonomous concept of the EU, the EU Member States need to take it into account when implementing and interpreting their copyright laws.[27] When it comes to NLP development, a reference must be made to the seminal *Infopaq* case, which suggests that copyright might subsist in a work comprising 11 consecutive words.[28] Advocate General Szpunar has suggested that even a book title such as *All Quiet on the Western Front* by Erich Maria Remarque naturally enjoyed copyright protection, together with the work as a whole.[29]

Related rights, such as those of performers, phonogram producers, *sui generis* database creators, broadcasters, and press publishers, are pertinent to NLP. Identifying the beneficiaries of these rights presents complexities akin to authorship determination. Particularly, when the beneficiary is a performer, verifying their identity is essential, highlighting similar challenges to identifying a work's author. This reflects the nuanced legal aspects of acknowledging and adhering to the rights relevant to NLP tasks.[30]

However, not every piece of language data is subject to copyright and the related rights protection. In the research literature, three types of language data could be outlined: works not covered by copyright (e.g. legal statutes, official documents), "safe" texts (e.g. manuals, technical documents, and official reports[31]), and

---

[25] Judgment of the Court of Justice of the European Union of 12 September 2019, case No. C-683/17, Cofemel Sociedade de Vestuário SA v. G-Star Raw CV, para. 29. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=217668&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=9984025 [last viewed 14.04.2024].

[26] Judgment of the Court of Justice of the European Union of 1 March 2012, case No. C-604/10, Football Dataco Ltd, et al. v. Yahoo! UK Ltd, et al. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=119904&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=2084171 [last viewed 14.04.2024].

[27] E.g, for further discussion on the evolution of the concept of work in Estonian copyright law, see *Kelli, A., Lepik, G.* Originality as a Key Concept of the Estonian Copyright Law. In: Handbook on Originality in Copyright. *Gupta, I.* (ed.). Singapore: Springer, 2023, pp. 1–19. Available: https://doi.org/10.1007/978-981-19-1144-6_10-1 [last viewed 14.04.2024].

[28] Judgment of the Court of Justice of the European Union of 16 July 2009, case No. C-5/08, Infopaq International A/S v. Danske Dagblades Forening, para. 48, 51. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=72482&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=1892287 [last viewed 14.04.2024].

[29] Opinion of Advocate General Szpunar delivered on 25 October 2018, case No. C-469/17, Funke Medien NRW GmbH v. Bundesrepublik Deutschland, para. 1. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62017CC0469&qid=1669110125285 [last viewed 14.04.2024].

[30] *Ilin, I.* Legal Regime of the Language Resources in the Context of the European Language Technology Development. In: Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2019. Lecture Notes in Computer Science, *Vetulani, Z., Paroubek, P., Kubis, M.* (eds). Springer, Cham, Vol. 13212, 2022, pp. 367–376. Available: https://doi.org/10.1007/978-3-031-05328-3_24 [last viewed 14.04.2024].

[31] As a matter of fact, in the Funke Medien case, CJEU asserted that military status reports "can be protected by copyright only if those reports are an intellectual creation of their author which reflect the author's personality and are expressed by free and creative choices made by that author in drafting those reports". Judgment of the Court of Justice of the European Union of 29 July 2019, case No. C-469/17, Funke Medien NRW GmbH v. Bundesrepublik Deutschland, para. 25. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=216545&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=10024671 [last viewed 14.04.2024].

copyright-protected texts.[32] Nonetheless, from a practical standpoint, relying solely on texts that are not covered by copyright to build datasets is insufficient. The limited volume and variety of such data would not support the development of an effective language model in that the use of "safe" and copyright-protected text becomes uninventable.[33]

To determine whether a text is protected by copyright, or if a "safe" text is copyrighted, the originality of the text must be assessed. The extensive body of caselaw harmonises the concept of originality at the EU level. It was already indicated in the *Infopaq* case that "copyright within the meaning of Article 2(a) of Directive 2001/29 is liable to apply only in relation to a subject-matter which is original in the sense that it is its author's own intellectual creation".[34] As seen, the originality is rooted in the concept of the author's creativity. In most EU Member States, copyright protection is contingent upon the work being a product of the author's intellect and personality, with the author typically being defined as a human being (e.g., Germany, Spain, France, Estonia). The human author approach is also supported by the analysis of Article 6*bis* of the Berne Convention, which regulates moral rights.[35]

The "human authorship" requirement introduces legal complexities for texts produced by AI with minimal human input. Copyright law analysis indicates that works generated without human effort are not eligible for copyright protection. Aware of this issue, AI service providers, including *OpenAI*, have instituted measures to curtail the use of AI-generated content for competing NLP development. Consequently, *OpenAI's* Terms of Use explicitly prohibit utilizing its outputs "to develop models that compete with *OpenAI*".[36] It is pointed out in the legal literature that since "the outputs of these applications are not protected by copyright, copyright exceptions, including the TDM exceptions, cannot apply to them".[37]

The act of gathering data not directly from the original sources but through intermediaries, such as social media platforms and repositories, introduces further legal intricacies. This approach imposes an additional layer of rights, typically those associated with *sui generis* database makers, necessitating careful legal navigation to address these rights adequately.[38]

---

[32] *Truyens, M., Van Eecke, P.* Legal aspects of text mining. Computer Law & Security Review, 30(2), 2014, pp. 153–170. Available: https://doi.org/10.1016/j.clsr.2014.01.009 [last viewed 14.04.2024].

[33] *Ilin, I., & Kelli, A.* The use of human voice and speech in language technologies: the EU and Russian intellectual property law perspectives. Juridica International, Vol. 28, 2019, pp. 17–27.

[34] Judgment of the Court of Justice of the European Union of 16 July 2009, case No. C-5/08, Infopaq International A/S v. Danske Dagblades Forening, para. 37. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=72482&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=1892287 [last viewed 14.04.2024].

[35] For further discussion, see *Kamocki, P., Bond, T., Lindén, K., Margoni, T., Kelli, A., Puksas, A.* Mind the Ownership Gap? Copyright in AI-generated Language Data, 2024. Linköping University Electronic Press (forthcoming).

[36] OpenAI. Terms of use. Effective: January 31, 2024. Available: https://openai.com/policies/terms-of-use [last viewed 14.04.2024].

[37] *Kamocki, P., Bond, T., Lindén, K., Margoni, T., Kelli, A., Puksas, A.* Mind the Ownership Gap? 2024 (forthcoming).

[38] *Kamocki, P., Hannesschläger, V., Hoorn, E., Kelli, A., Kupietz, M., Lindén, K., Puksas, A.* Legal Issues Related to the Use of Twitter Data in Language Research, 2022, pp. 68–75. In: Selected Papers from the CLARIN Annual Conference 2021. Linköping Electronic Conference Proceedings, *Monachini, M., Eskevichm, M.* (eds). Available: https://doi.org/10.3384/ecp1897 [last viewed 14.04.2024].

Therefore, the challenge of data access emerges distinctly. Recent legal actions in the US against OpenAI[39] underscore the conflict between authors or rights holders, who demand recognition and fair compensation for their copyrighted works, and NLP developers, who need broad access to data to build efficient language models. Nonetheless, the issue of accessing language data is not simply about balancing copyright and related rights against NLP developers' needs; it encompasses more intricate complexities.

NLP technology has substantial economic and social implications, engaging governmental and individual stakeholders in its advancement. End-users of NLP applications seek to safeguard their fundamental rights (e.g. the right to privacy), whereas governments focus on promoting economic, cultural, and social advancement. Balancing copyright and related rights with both public and private interests is crucial. Yet, this balancing act becomes increasingly challenging in the digital era, where the distinction between public and private interests often blurs. Interests among governments, businesses, and individuals are dynamic and may conflict, complicating the effort to reconcile these competing priorities.[40]

Even without exhaustive analysis, it is apparent that copyright and/or related rights typically safeguard data requisite for NLP development. Furthermore, additional restrictions may arise from the terms of service of repositories or social media platforms from which the data is harvested, along with an extra tier of rights, such as *sui generis* database rights. Consequently, the principal legal challenges in NLP development revolve around the accurate identification of the appropriate legal foundations for such activities.

Copyright-wise, NLP development mainly operates within two frameworks: the contractual model and the exception model. Additionally, hybrid models can arise, mixing contractual agreements with copyright exceptions for language data use. Due to its legal complexities and challenges, the following sections will examine both models, especially the exception model. This analysis is vital for grasping the complicated legal environment of NLP development and managing the delicate equilibrium between utilizing data for technological progress and adhering to copyright and related rights.

2.2. NLP development under the contractual model

The political discussions surrounding contractual models are not novel. The "Licences for Europe" stakeholder dialogue, initiated in 2013, includes Working Group (WG) 4, which outlines the Commission's objective as: "to promote the efficient use of text and data mining (TDM) for scientific research purposes. TDM currently requires contractual agreements between users (e.g. typically research institutions)

[39] Class Action Complaint and Demand for Jury Trial in case No. 1:24-cv-00084, Nicholas Gage v. Microsoft, OpenAI, United States District Court for the Southern District of New York. Available: https://fingfx.thomsonreuters.com/gfx/legaldocs/klvydkdklpg/OPENAI%20COPYRIGHT%20LAWSUIT%20basbanescomplaint.pdf [last viewed 14.04.2024]; Complaint and Demand for Jury Trial in case No. 1:23-cv-11195, the New York Times company v. Microsoft, OpenAI, United States District Court for the Southern District of New York. Available: https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf [last viewed 14.04.2024].

[40] For instance, some individuals may argue for adequate representation in training data to reduce potential discrimination, while authors might demand fair compensation for data use, possibly decreasing available training data. This illustrates the challenge of balancing ethical considerations in AI development with authors' rights, highlighting the complexity of aligning various interests in NLP technology.

and rights holders (e.g. publishers of scientific journals) to establish the modalities for technical access to the relevant data sets".[41]

The contractual model requires AI developers to obtain copyright holders' permission to use copyrighted materials, favouring rightsholders to ensure their participation in the AI-driven value chain. This motivates them to publicize agreements to influence AI development regulations. This model fosters collaboration between copyright owners and AI developers, supporting innovation within legal boundaries. For example, a press release highlights a strategic partnership between *BRIA*, a proprietary AI visual content tool developer, and Getty Images, a leading global visual content creator and marketplace. This agreement permits creatives to adapt images to their specific requirements using intuitive AI tools on *Getty Images*' platform.[42] In contrast, there is a legal case where *Getty Images* initiated a lawsuit against *Stability AI*, the creator of the AI art generator *Stable Diffusion*, accusing it of infringing upon *Getty Images*' rights.[43]

The contractual model tackles legal challenges in commercial research and creating language datasets and models for business. Yet, it faces practical difficulties, such as identifying rightsholders for anonymous blog posts or orphan works, leading to time-consuming and costly processes that can slow NLP development. The need for vast amounts of language data exacerbates this issue.

The use of the contractual model differs among internet giants like *Google* and *Yandex* in NLP development. For example, *Yandex's* voice assistant Alice sources input not just from its app but also from *Yandex's* other services (like navigation, taxi, and translation), integrating appropriate clauses in the licenses for these services.[44] *OpenAI's ChatGPT* terms stipulate that *OpenAI* can use content (input and output data) to deliver, maintain, develop, and enhance its services.[45] However, when developers lack proprietary materials for language data, employing the contractual model becomes expensive and time-consuming.

The contractual model can be broadly viewed as a social contract between creators and AI developers, suggesting a mutual understanding and agreement on the use of creative content in AI development.

Research literature suggests that generative AI systems have the potential to replace human creators in certain contexts. To address the potential displacement of human creators by generative AI systems, a proposed solution is to "introduce an output-oriented levy system that imposes a general payment obligation on all providers of generative AI systems in the EU. In contrast to remuneration systems based on

[41] European Commission. "Licences for Europe" stakeholder dialogue. 22 December 2017. Available: https://digital-strategy.ec.europa.eu/en/library/licences-europe-stakeholder-dialogue [last viewed 14.04.2024].

[42] BRIA Partners with Getty Images to Transform Visual Content Through Responsible AI. 25 October 2022. Available: https://investors.gettyimages.com/news-releases/news-release-details/bria-partners-getty-images-transform-visual-content-through [last viewed 14.04.2024].

[43] *Vincent, J.* Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement. 6 February 2023. Available: https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion [last viewed 14.04.2024].

[44] *Ilin, I.* Legal Regime of the Language Resources in the Context of the European Language Technology Development. In: Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2019. Lecture Notes in Computer Science, *Vetulani, Z., Paroubek, P., Kubis, M.* (eds). Springer, Cham: 2022, Vol. 13212, pp. 367–376. Available: https://doi.org/10.1007/978-3-031-05328-3_24 [last viewed 14.04.2024].

[45] OpenAI. Terms of use. Effective: January 31, 2024. Available: https://openai.com/policies/terms-of-use [last viewed 14.04.2024].

AI training activities, this alternative approach would not weaken the position of the European AI sector or make the EU less attractive as a region for AI development. Even more importantly, an output-oriented AI levy system can be combined with mandatory collective rights management".[46]

The authors regard the proposal for an output-oriented levy system as innovative and stimulating, recognizing its potential benefits. However, implementing such an approach would involve significant political decision-making on many practical aspects. Essentially, this model would shift from granting authors exclusive rights (the right to prevent others from using their work for AI development) to entitling them to remuneration. This shift raises questions about how to distribute the collected levy among rightsholders (blank tape or private copying levy system serve here as a possible example) and to what extent AI developers must disclose the works and related rights objects utilized. Although the AI Act mandates transparency obligations, the question remains as to how detailed such transparency reports must be and whether this level of detail is technically feasible. Consequently, the likelihood of this suggested model being adopted appears slim.

In summary, the contractual model for accessing copyrighted materials requires identifying rightsholders (including the reliance on the extended collective rights management) and establishing negotiation and licensing agreements, adding complexity and potential costs to NLP development. This challenge is intensified by the requirement for a large volume of works for AI development. The proposal to implement an output-oriented levy system is innovative but encounters a need for political decisions and numerous practical hurdles.

2.3. NLP development under the exception model

The authors explore other copyright exceptions relevant to NLP development before delving into the TDM exception. The authors contend that aside from the TDM exception, NLP development benefits from the exception for temporary acts of reproduction, the personal use exception, the quotation right and the research exception established prior to the adoption of the DSM Directive.[47]

The key copyright exceptions pertinent to NLP development are outlined in Article 5 of the InfoSoc Directive. Case law consistently reiterates that exceptions and limitations to the reproduction right and the right of communication to the public are exhaustively enumerated in Article 5 of the InfoSoc Directive.[48] This does not imply that other directives are precluded from introducing copyright exceptions and

[46] *Senftleben, M.* Generative AI and Author Remuneration. International Review of Intellectual Property and Competition Law, 54, 2023, pp. 1535–1560. Available: https://doi.org/10.1007/s40319-023-01399-4 [last viewed 14.04.2024].

[47] For an in-depth discussion on the exceptions and their impact in the area before the DSM Directive's implementation, see *Eckart de Castilho, R., Dore, G., Margoni, T., Labropoulou, P. & Gurevych, I.* A Legal Perspective on Training Models for Natural Language Processing. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, ELRA, 2018, pp. 1267–1274. Available: http://www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf [last viewed 14.04.2024]; *Kelli, A., Tavast, A., Lindén, K., Vider, K., Birštonas, R., Labropoulou, P., Kull, I., Tavits, G., Värv, A., Stranák, P., Hajic, J.* The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies. In: Selected Papers from the CLARIN Annual Conference 2019, *Simov, K., Eskevich, M.* (eds). Linköping University Electronic Press, 2020, pp. 53–65. Available: http://doi.org/10.3384/ecp2020172008 [last viewed 14.04.2024].

[48] Judgment of the Court of Justice of the European Union of 16 November 2016, case No. C-301/15, Marc Soulier, Sara Doke v. Premier ministre, Ministre de la Culture et de la Communication, para. 34. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=185423&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=164343 [last viewed 14.04.2024].

limitations. Instead, the concept is that exceptions and limitations should not derive from broad general principles. Nevertheless, these general principles (e.g., freedom of speech) can play a role in influencing the interpretation of copyright exceptions and limitations as stipulated by directives.

The InfoSoc Directive outlines mandatory exceptions, which EU Member States are required to implement, and optional exceptions, which Member States can choose to adopt or not, relevant to NLP development. The optional nature of several exceptions has led to fragmentation within the regulatory frameworks governing TDM in the EU prior to the DSM Directive. It remains unclear whether the applicability of the referenced copyright exceptions can be contractually restricted. Given the policy objectives underlying these exceptions, it could be argued that they possess a mandatory character, rendering any contractual terms that limit their application void.

The sole mandatory exception pertains to temporary acts of reproduction, as delineated in Article 5(1) of the InfoSoc Directive. This provision holds particular relevance to NLP development. Recital 9 of the DSM Directive notes that TDM activities not involving reproduction, or where reproductions are covered by this exception, should continue to be permissible. This facilitates certain NLP processes without infringing copyright laws, provided they do not exceed the scope of this exception.

The personal use exception[49] might benefit certain NLP developments, yet it comes with notable restrictions. Firstly, it cannot underpin large-scale activities. Secondly, it is unclear whether a language model developed under this exception can be utilized for other purposes, such as business.

Article 5(3)(d) of the InfoSoc Directive permits quotations for purposes like criticism or review if they pertain to a lawfully publicised work or subject matter. It stipulates that the source, including the author's name, should be cited unless impossible, and the use must align with fair practice and be proportionate to the intended purpose. The right to quotation, permitting the use of short excerpts from a work, could be considered a legal foundation for NLP development. However, EU case law[50] clarifies that quoting presupposes an intention to engage in a "dialogue" with the work. Given that NLP development does not entail such dialogue, the right to quotation does not provide a suitable legal basis for it.

Before adopting the TDM exception in the DSM Directive, the research exception outlined in Article 5(3)(a) of the InfoSoc Directive likely served as a legal basis for NLP development. This exception permits "scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved".

All exceptions must align with the three-step test. Internationally, this test is embedded within the Berne Convention (Article 9(2)), the TRIPS Agreement (Article 13), and the WIPO Copyright Treaty[51] (Article 10). The three-step test is articulated at the EU level in Article 5(5) of the InfoSoc Directive. The CJEU, drawing upon established jurisprudence, maintains that the exceptions aim to establish a "fair

[49] InfoSoc Directive Art. 5 (2) (b).

[50] Judgment of the Court of Justice of the European Union of 29 July 2019, case No. C-476/17, Pelham GmbH, Moses Pelham, Martin Haas v. Ralf Hütter, Florian Schneider-Esleben, para. 71. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=216552&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=547686 [last viewed 14.04.2024].

[51] WIPO Copyright Treaty. Adopted in Geneva on December 20, 1996. Available: https://www.wipo.int/wipolex/en/text/295166#P86_11560 [last viewed 14.04.2024].

balance" between the rights and interests of authors, on the one hand, and the rights of users pertaining to protected subject matter, on the other.[52]

2.4. NLP development under the TDM exception

Text and data mining (TDM) is "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".[53] This legal definition offers a broad understanding of TDM activities, emphasizing their primary objective of generating new information. Although this definition provides a comprehensive framework, TDM activities involve a variety of techniques tailored to specific mining purposes, which may lead to potential legal complexities across different fields. Although the law groups "text" and "data" mining together, their technical processes may not necessarily be identical. While many researchers consider text mining a subset of data mining[54], the distinction lies in the sources they utilize to achieve their objectives. Data mining techniques draw upon diverse datasets, such as spatial data, network data, DNA sequence data, multimedia, and stream data, tailored to the specific objectives of the TDM activity.[55] Conversely, text mining focuses on narrower sources and predominantly relies on text analysis, serving as a fundamental activity for NLP development.

The DSM Directive introduced two TDM exceptions: the general-purpose TDM exception (Article 4) and the TDM exception for research (Article 3). In this article, these exceptions are collectively referred to as the TDM exceptions.

Text mining techniques are crucial for constructing language models. These techniques may rely on corpora, such as the *Universal Dependencies* treebanks[56] and the *Common Crawl* dataset[57], which are created through the collection, copying, structuring, and labelling of language data stored in various formats. Simultaneously, it remains challenging to directly align the scope of the TDM exceptions with the specific stage of NLP development outlined above. While certain activities within the NLP development stages may reasonably fall under the TDM exception, it is important to recognize that the entirety of the process cannot be exclusively accommodated by this exception.

The TDM exceptions comprise a myriad of legal intricacies. However, space limitations require a concentrated examination, which revolves around several key aspects: remuneration to rightsholders for the utilization of copyrighted works and related rights objects, the scope of the exception, the lawful source prerequisite, and the legal standing of models developed under the TDM exception for scientific research. These facets are chosen due to their integration and perceived significance in addressing the challenges inherent to the TDM exception.

---

[52] Judgment of the Court of Justice of the European Union of 3 September 2014, case No. C-201/13, Johan Deckmyn, et al. v. Helena Vandersteen et al., para. 26. Available: https://curia.europa.eu/juris/document/document.jsf?mode=DOC&pageIndex=0&docid=157281&part=1&doclang=EN&text=&dir=&occ=first&cid=689582 [last viewed 14.04.2024].

[53] Article 2(2) of the DSM Directive.

[54] *Zong, C., Xia, R., & Zhang, J.* Text data mining. Vol. 711, Singapore: Springer, 2021, p. 712. Available: https://doi.org/10.1007/978-981-16-0100-2 [last viewed 14.04.2024].

[55] *Han, J., Pei, J., & Tong, H.* Data mining: Concepts and Techniques. 4th edition. Morgan Kaufmann: 2022, p. 752.

[56] Universal Dependencies (UD), version 2. Available: https://universaldependencies.org/ [last viewed 14.04.2024].

[57] Common Crawl repository. Available: http://commoncrawl.org/ [last viewed 14.04.2024].

Recital 17 of the DSM Directive clarifies that implementing the TDM exception for research purposes results in negligible harm to rightsholders and, therefore, does not require compensation. The scenario differs concerning the general-purpose TDM exception, as its regulation includes the opt-out right, enabling the exclusion of TDM activities.[58] Essentially, the opt-out right affords rightsholders the opportunity to obtain remuneration for TDM activities. The Artificial Intelligence Act (AI Act) also bolsters the potential right to remuneration. Article 53(1)(c) obliges providers[59] of general-purpose AI models[60] to establish a policy for adhering to Union copyright law, specifically to recognize and adhere to the opt-out right as delineated in Article 4(3) of the DSM Directive. Recital 107 of the AI Act elucidates that to augment transparency regarding the data utilized in the pre-training and training of general-purpose AI models, encompassing copyrighted text and data, providers of such models are required to draft and publicly release a sufficiently detailed summary of the content utilized for training purposes. This provision enables rightsholders to assert and uphold their rights. However, uncertainty persists regarding how the opt-out right can be exercised. For instance, is it sufficient for collective management organizations to announce that they do not permit the use of the copyrighted content of the rightsholders they represent? An adequate standard remains to be developed and tested in court.

It can be argued that, when Article 4(3) of the DSM Directive is combined with the transparency obligation established by the AI Act, it allows rightsholders to claim remuneration for the utilization of their copyrighted content under the general-purpose TDM exception. This could potentially disadvantage EU-based AI companies if the corresponding regulatory framework in other jurisdictions, such as the US and China, is structured differently – for instance, if it does not afford a remuneration right.

When analysing the scope of the TDM exceptions, it is evident that these exceptions primarily restrict the reproduction right and the right to make extractions. Nonetheless, TDM activities may also encompass the addition of annotations to data. The authors argue that the TDM exceptions should be construed broadly to permit such annotations and other adoptions dictated by technical necessity.

The TDM exceptions specify lawful access to data as a prerequisite for their application. The concept of lawful access is not as straightforward as it might initially appear. Recital 10 of the DSM Directive references instances where researchers have lawful access to content, for example, through subscriptions to publications or open access licenses. However, the terms of these licenses could explicitly exclude TDM activities. Article 7(1) of the DSM Directive stipulates that any contractual provision contrary to the TDM exception for research is deemed unenforceable. Recital 14 of the DSM Directive states that lawful access should also encompass access to content

[58] See, DSM Directive Article 4 (3).

[59] Article 3(3) of the AI Act defines "provider" as "a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge".

[60] Article 3(63) of the AI Act defines "general-purpose AI model" as "an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are released on the market".

that is freely available online. This raises the question of whether lawful access is possible for works communicated to the public without the rightsholder's consent. It is proposed in scholarly literature that "the requirement of lawful access should only cover the behaviour of the beneficiary of the exception and not extend to the status of the accessed source".[61] This approach could potentially facilitate TDM for NLP development and help avoid many legal uncertainties and practical problems. However, there are concerns regarding the compatibility of this approach with the three-step test. Lawful access typically necessitates a lawful source.

The final issue addressed in this section concerns the legal standing of models developed under the TDM exception for scientific research. One of the primary legal challenges revolves around the interpretation of what constitutes "scientific research", and which entities are eligible to perform TDM under this exemption. Ensuring fair access to language data for academic and commercial entities is vital for fostering fair competition and advancing progress in NLP research and applications. However, the lack of clarity surrounding this issue can hinder the full engagement of commercial entities and non-traditional research institutions in NLP development as they seek to avoid potential copyright infringement risks. Specifically, the question arises whether models created based on the TDM exception for research (Article 3 of the DSM Directive) can be utilized for business purposes. Recital (11) of the DSM Directive emphasises that "research organisations should also benefit from such an exception when their research activities are carried out in the framework of public-private partnerships".

Despite the policy objective of supporting public-private partnerships[62], it is essential to consider the entire regulatory framework surrounding TDM and model development. Additionally, the AI Act requires providers of general-purpose AI to enhance transparency regarding the use of copyright-protected data.[63] This enables rightsholders to monitor the usage of their copyrighted material and enforce their rights effectively.

If it emerges that content subject to an opt-out right was used by research organizations under the exception for text and data mining intended for research, rendering the opt-out right unenforceable, then the subsequent use of such a model by a private company could potentially conflict with copyright laws, such as the three-step test.

3. Legal challenges related to the output of NLP

3.1. Originality of prompts under copyright

The relationship between AI-generated content and prompts is fundamental to the operation of AI systems and their output generation. Prompts function as input cues that guide the content generation process, whether by providing explicit instructions, influencing the learning process through training data, or conditioning the generation of content based on contextual factors. Understanding this relationship

[61] *Margoni, T.* Saving research: Lawful access to unlawful sources under Art. 3 CDSM Directive? Kluwer Copyright Blog, 2023. Available: https://copyrightblog.kluweriplaw.com/2023/12/22/saving-research-lawful-access-to-unlawful-sources-under-art-3-cdsm-directive/ [last viewed 14.04.2024].

[62] For further discussion on the academia-industry cooperation, see *Kelli, A., Mets, T., Jonsson, L., Pisuke, H., Adamsoo, R.* The changing approach in Academia-Industry collaboration: From profit orientation to innovation support. Trames Journal of the Humanities and Social Sciences, 17(3), 2013, pp. 215–241. Available: http://doi.org/10.3176/tr.2013.3.02 [last viewed 14.04.2024].

[63] Recital 107 of the AI Act.

is essential for effectively using and interpreting AI-generated content across different applications. However, the legal status of prompts remains ambiguous. This issue is increasingly significant due to specialised marketplaces' active distribution of prompts today.[64]

Although prompts may be viewed as instructions for AI, whether they qualify for copyright protection remains uncertain. A core principle of copyright law is that protection is granted solely to works considered "original". Within the EU *acquis*, the criterion for copyright protection hinges on whether a work can be seen as the author's "own intellectual creation". The Court of Justice of the European Union (CJEU) has asserted that an original work results from intellectual creation. The author expresses his creative ability in an original manner by making free and creative choices such that the resulting shape reflects his personality.[65] At the same time, facts, ideas and utilitarian processes are excluded from copyright protection.[66] In that, the question of whether prompts are original under copyright law involves a nuanced analysis of creativity and human involvement.

While some prompts may be straightforward and purely functional, others may involve creative choices, linguistic nuances, or artistic elements. For example, prompts used in storytelling AI models may include specific character descriptions, plot outlines, or dialogue prompts that reflect creative input.

Another factor to consider is the degree of human involvement in creating prompts. If prompts are generated entirely by AI systems without human intervention, no copyright subsists in them.

An intriguing issue arises regarding whether the use of numerous prompts as input to an AI system could result in joint authorship with the AI. In the case concerning *Théâtre D'opéra Spatial*, the applicant explained that he "input numerous revisions and text prompts at least 624 times to arrive at the initial version of the image". The United States Copyright Office Review Board did not uphold the claim. The Board acknowledges that prompting can involve creativity and that some prompts may be protected as literary works. However, this does not mean that providing text prompts to *Midjourney* forms the generated images.[67] The authors argued that the issue of joint authorship with AI has not yet been resolved.

There is also a lack of consensus in addressing prompts from a contractual standpoint. For example, the terms of use (ToS) on *Prompt Marketplace* describe prompts as intellectual property (IP) objects without providing additional details while maintaining ownership rights for their creators (authors).[68] Furthermore, the marketplace imposes quality standards for prompts. As outlined in its Prompt

[64] E.g., Prompt Marketplace. Available: https://promptbase.com [last viewed 14.04.2024]; AI Prompt Marketplace. Available: https://promptrr.io [last viewed 14.04.2024]; AIFrog. Available: https://www.aifrog.io [last viewed 14.04.2024].

[65] Judgment of the Court of Justice of the European Union of 11 June 2020, case No. C-833/18, SI, Brompton Bicycle Ltd v. Chedech/Get2Get. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=227305&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=1727690 [last viewed 14.04.2024].

[66] Article 9(2) TRIPS Agreement.

[67] U.S. Copyright Office, Letter 05.09.2023, Re: Second Request for Reconsideration for Refusal to Register Théâtre D'opéra Spatial (SR # 1-11743923581). Available: https://www.copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf [last viewed 14.04.2024].

[68] PromptBase. Terms of Service. Available: https://promptbase.com/tandcs [last viewed 14.04.2024].

Submission Guidelines,[69] prompts are expected to offer value to buyers,[70] be original, and not overly simplistic or easily predictable. It is plausible that these quality requirements be utilized to assess prompt originality under copyright law. However, not all prompt marketplaces prioritize addressing the legal status of prompts from an intellectual property law perspective. The terms of service (ToS) of the AI Prompt Marketplace describe prompts as the digital content provided by the seller.[71] In this scenario, buyers acquire a license to use the prompts rather than ownership. This transition from creator to seller ownership raises questions about the basis of ownership and the seller's authority to grant usage licenses. Additionally, although the Digital Content Directive[72] does not directly address intellectual property issues, it implies that the digital content provided should not infringe upon any third-party rights, including intellectual property rights.[73] This implies that if prompts are subject to copyright protection, the seller must obtain appropriate authorization and uphold the author's moral rights. This task may pose challenges, especially when purchasing prompts through online marketplaces.

3.2. Legal framework for content generated by NLP

As a rule, content generated by NLP gives rise to the same issues as AI-generated content. Due to space limitations, not all aspects of NLP-generated content can be thoroughly addressed. Therefore, the article briefly touches upon ethical considerations and authorship issues.

AI-generated (including NLP-generated) content is becoming more common, but we need to figure out how to tell if it's made by humans or AI. This is applicable to different areas, including journalism, art, research and education. For example, fake news made by AI can be harmful, AI art raises questions about who made it, and using NLP content in schoolwork can be seen as academic fraud.[74] One could contend that AI applications, such as NLP, have already significantly impacted society. For example, a *Reuters* article highlighted that there exists a multitude of books where *ChatGPT* is credited as either an author or a co-author, underscoring the profound influence of AI on contemporary authorship and literary creation.[75] Similarly, AI-driven "trading

---

[69] Prompt Submission Guidelines, PromptBase. Available: https://promptbase.com/prompt-guidelines [last viewed 14.04.2024].

[70] The principle of "if value then right" suggests that individuals or entities should only possess rights to something if they've added value to it or if the thing itself is inherently valuable. Works that need minimal human input are often considered less valuable, thus requiring minimal copyright protection. For further discussion, see *Dreyfuss, R. C.* Expressive genericity: trademarks as language in the Pepsi generation. Notre Dame Law Review, No. 65, 1989, p. 397 and *Lemley, M. A.* How Generative AI Turns Copyright Upside Down, 2023, pp. 12–13. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4517702 [last viewed 14.04.2024].

[71] Terms of service, AI Prompt Marketplace. Available: https://promptrr.io/terms-of-service/ [last viewed 14.04.2024].

[72] Directive (EU) 2019/770 of the European Parliament and of the Council on certain aspects concerning contracts for the supply of digital content and digital services (Digital Content Directive) [2019] OJ L136/1. Available: https://eur-lex.europa.eu/eli/dir/2019/770/oj [last viewed 14.04.2024].

[73] Recital 54, Article 7 of the Digital Content Directive.

[74] *Cotton, D. R. E., Cotton, P. A., & Shipway, J. R.* Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 2024, pp. 228–239. Available: https://doi.org/10.1080/14703297.2023.2190148 [last viewed 14.04.2024].

[75] *Bensinger, G.* Focus: ChatGPT launches boom in AI-written e-books on Amazon. 21 February 2023. Available: https://www.reuters.com/technology/chatgpt-launches-boom-ai-written-e-books-amazon-2023-02-21/ [last viewed 14.04.2024].

bots" have been deployed to supplant human roles in financial markets,[76] illustrating further instances of AI's expanding footprint across various sectors. This suggests that the advent of AI may render traditional roles obsolete, potentially displacing authors from the market. Consequently, legal scholars are actively seeking solutions to address these emerging ethical challenges and technological unemployment. This could serve as an additional argument to support the imposition of a general payment obligation on all providers of generative AI systems in the EU, as suggested by Senftleben.[77] The use of AI is not inherently unethical. Adopting a Luddite ideology as the primary ethical framework for AI use is problematic, as it would imply that even spell-checkers should be prohibited to enhance "real" creativity. Given the challenges in determining whether content is AI-generated, transparency regarding the role of AI in content generation is essential and should be established as a fundamental starting point. It is suggested in legal literature that there should be a mandatory legal obligation to declare if content is AI-generated.[78]

The topic of AI-generated content is continuously relevant. Research literature distinguishes between AI-assisted output and AI-generated output, with the latter typically falling outside copyright protection.[79] We are interested in the latter.

AI systems can independently generate content, triggering discussions regarding the ownership of such creations. In some cases, the ownership of AI-generated content may be attributed to the human creators who developed or trained the AI system, or to the rightsholders of works used to train the AI system. However, this may not always be straightforward, especially as AI systems become more advanced and independent (autonomous) in their decision-making processes.

In response to the growing use of AI in creative processes, the US Copyright Office (USCO) issued a notice of inquiry (NOI)[80] seeking input on various aspects related to AI-generated content. These include questions concerning ownership rights, transparency requirements, and the legal status of AI-generated outputs. The need for clarity in copyright registration for AI-generated works became apparent in the case of "Zarya of the Dawn", where the USCO cancelled portions of AI-generated artwork from the copyright registration.[81] This decision underscored the importance of establishing clear guidelines for registering AI-generated content and ensuring transparency in the copyright registration process. Similarly, in the case

[76] *Bloom, J.* Could AI 'trading bots' transform the world of investing? 1 February 2024. Available: https://www.bbc.com/news/business-68092814 [last viewed 14.04.2024].

[77] *Senftleben, M.* Generative AI and Author Remuneration. International Review of Intellectual Property and Competition Law, 54, 2023, pp. 1535–1560. Available: https://doi.org/10.1007/s40319-023-01399-4 [last viewed 14.04.2024].

[78] *Kamocki, P., Bond, T., Lindén, K., Margoni, T., Kelli, A., Puksas, A.* Mind the Ownership Gap? 2024 (forthcoming).

[79] It is argued in the literature that "AI-assisted output to qualify as a protected "work": the output is (1) in relation to "production in the literary, scientific or artistic domain"; (2) the product of human intellectual effort; and (3) the result of creative choices that are (4) "expressed" in the output". *Hugenholtz, P. B., Quintais, J. P.* Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output? IIC 52, 2021, p. 1212. Available: https://doi.org/10.1007/s40319-021-01115-0 [last viewed 14.04.2024].

[80] Notice of inquiry and request for comments. U.S. Copyright Office, Library of Congress. Federal Register, Vol. 88, No. 167, Wednesday, August 30, 2023. Available: https://www.govinfo.gov/content/pkg/FR-2023-08-30/pdf/2023-18624.pdf [last viewed 14.04.2024].

[81] U.S. Copyright Office, Letter 21.02.2023, Re: Zarya of the Dawn (Registration #VAu001480196). Available: https://www.copyright.gov/docs/zarya-of-the-dawn.pdf [last viewed 14.04.2024].

of *Théâtre D'opéra Spatial* by Jason Allen[82], the applicant encountered challenges when attempting to register a work generated by the AI system *Midjourney*.[83] Despite the applicant's arguments about creative input, the USCO refused registration, emphasizing the requirement of human authorship and the necessity of disclosing AI-generated content.

The decision not to register AI-generated content due to the absence of human authorship underscores the complexities of determining ownership and copyright eligibility for such works. Meanwhile, an alternative avenue for resolving ownership issues is through contractual agreements. When users engage with AI services or platforms, they typically consent to the terms of service or end-user license agreements presented by the platform. These agreements often contain clauses that define the ownership of content produced using the AI service. For example, in *OpenAI's* Terms of Service[84], users retain ownership of the content they generate with *OpenAI's* services, while *OpenAI* may utilize this content for various purposes, including service enhancement. Users also have the option to decline *OpenAI's* use of their content to train its models.[85] Similarly, in *NotionAI's* Terms of Service, users retain ownership of the content they generate on the *NotionAI* platform. *NotionAI* respects users' rights to their content while leveraging it to improve their services.[86] However, users may have limited leverage when negotiating the terms of service provided by AI service providers. Furthermore, the absence of legal precedent in this area makes it challenging to anticipate how these agreements would function in practice.

In summary, navigating copyright ownership in the context of AI-generated content requires careful consideration of legal principles, technological capabilities, and ethical implications. Balancing the interests of human creators, AI entities, and other stakeholders is essential for developing robust and equitable frameworks that promote innovation while respecting rights and responsibilities.

Summary

The saying "US innovates, China replicates, and EU regulates" contains some truth. Europe's AI sector competes globally, and its stringent regulations, while ensuring a high level of IP protection, ethical standards and data protection, may not automatically translate into competitiveness. The lack of legal clarity in EU regulations introduces complexities, potentially hindering innovation in the EU compared to regions with more lenient policies.

The challenge for Europe lies in balancing rigorous standards with maintaining a competitive edge in the global AI landscape, ensuring it can innovate effectively while adhering to its values. This balance is crucial for Europe to remain a significant player in the international AI arena.

---

[82] U.S. Copyright Office, Letter 05.09.2023, Re: Second Request for Reconsideration for Refusal to Register Théâtre D'opéra Spatial (SR # 1-11743923581). Available: https://www.copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf [last viewed 14.04.2024].

[83] Midjourney platform. Available: https://www.midjourney.com/home [last viewed 14.04.2024].

[84] OpenAI. Terms of use. Effective: January 31, 2024. Available: https://openai.com/policies/terms-of-use [last viewed 14.04.2024].

[85] OpenAI's right to utilize the user data potentially creates privacy concerns and trade secret safety.

[86] Notion AI Supplementary Terms. Available: https://www.notion.so/notion/Notion-AI-Supplementary-Terms-fa9034c8b5a04818a6baf3eac2adddbb [last viewed 14.04.2024].

Language data often includes copyrighted works and objects of related rights (performances, phonograms, excerpts from databases, etc.), posing significant challenges for its use in NLP development. Methods like web scraping and data sharing bring up contract violations and copyright infringement issues. This situation requires a nuanced equilibrium between respecting authors' rights and meeting the developmental needs of NLP practitioners.

There are two primary models for developing NLP applications: the contractual model and the exception model. Additionally, a hybrid model incorporating elements from both can be identified. Each model has its own advantages and disadvantages.

The contractual model offers potential solutions for challenges arising in commercial research contexts. The challenge with the contractual model for NLP development is not primarily about AI developers' reluctance to share profits – though that may often be the case – but more about the difficulty in identifying rightsholders or their sheer volume, given the extensive language data required. This makes it nearly impossible to negotiate agreements with all involved. Unlike the music industry, which benefits from established collective management organizations for licensing, no equivalent structures exist for licensing language data. Hence, the contractual model cannot be the sole path forward; alternative approaches are also necessary.

The exception model refers to NLP development based on copyright exceptions. Prior to the DSM Directive era, NLP development relied on exceptions provided in the InfoSoc Directive, such as exceptions for temporary reproduction, personal use, quotation, and research. The problem is that, except for the temporary reproduction exception (InfoSoc Art. 5(1)), other exceptions are optional for EU Member States, leading to fragmentation within the EU. Despite this, these exceptions remain relevant for NLP development.

With the DSM Directive, two TDM exceptions were adopted: the general TDM exception and the TDM exception for research purposes. Both TDM exceptions face legal uncertainties concerning the concept of lawful access, specifically whether good faith access from an illegal source is permissible as analysed by Margoni.[87] The authors are of the opinion that the three-step test likely restricts this possibility.

The primary challenge of the general TDM exception is the opt-out right, which allows rightsholders to explicitly reserve the use of their content, forbidding its use for TDM. Firstly, there are technical issues, such as determining how to make such a reservation in a manner compatible with Article 4(3) of the DSM Directive. Secondly, the very existence of the opt-out right gives rise to several ambiguities. It would have been more straightforward to state that rightsholders are entitled to remuneration and to establish a remuneration system accordingly.

Due to the opt-out right, reinforced by the transparency obligations in the AI Act, the development and use of NLP applications outside academia is shifting towards the contractual model. The contractual model of NLP incurs costs and uncertainties for developers. One way forward to enhance NLP use and development in the EU would be to allow NLP applications developed under the TDM exception for research to be used for commercial purposes. However, the issue is unclear, and it is most likely not permissible.

Bearing in mind the *Infopaq* case (where 11 consecutive words could constitute a copyrighted work) and the approach of the US Copyright Office, it can be argued

[87] *Margoni, T.* Saving research: Lawful access to unlawful sources under Art. 3 CDSM Directive? Kluwer Copyright Blog, 2023. Available: https://copyrightblog.kluweriplaw.com/2023/12/22/saving-research-lawful-access-to-unlawful-sources-under-art-3-cdsm-directive/ [last viewed 14.04.2024].

that, depending on their nature, some prompts are protected by copyright. The main issue is whether the creation and input of prompts could lead to joint authorship of NLP-generated content. The current theoretical framework and legal practice does not answer the question clearly.

The AI-generated output is currently considered outside the scope of copyright protection. Therefore, its legal status is primarily regulated by the AI service provider's terms of service.

References

Bibliography

Barthélemy, F., Ghesquière, N., Loozen, N. et al. Natural language processing for public services. European Commission, Directorate-General for Digital Services. Publications Office of the European Union, 2022. Available: https://data.europa.eu/doi/10.2799/304724 [last viewed 14.04.2024].

Bensinger, G. Focus: ChatGPT launches boom in AI-written e-books on Amazon. 21 February 2023. Available: https://www.reuters.com/technology/chatgpt-launches-boom-ai-written-e-books-amazon-2023-02-21/ [last viewed 14.04.2024].

Bloom, J. Could AI 'trading bots' transform the world of investing? 1 February 2024. Available: https://www.bbc.com/news/business-68092814 [last viewed 14.04.2024].

BRIA Partners with Getty Images to Transform Visual Content Through Responsible AI. 25 October 2022. Available: https://investors.gettyimages.com/news-releases/news-release-details/bria-partners-getty-images-transform-visual-content-through [last viewed 14.04.2024].

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 2024, pp. 228–239. Available: https://doi.org/10.1080/14703297.2023.2190148 [last viewed 14.04.2024].

Dash, N. S., & Arulmozi, S. History, features, and typology of language corpora. Singapore: Springer, 2018. Available: https://doi.org/10.1007/978-981-10-7458-5 [last viewed 14.04.2024].

Dreyfuss, R. C. Expressive genericity: trademarks as language in the Pepsi generation. Notre Dame Law Review, No. 65, 1989.

Eckart de Castilho, R., Dore, G., Margoni, T., Labropoulou, P. & Gurevych, I. A Legal Perspective on Training Models for Natural Language Processing. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, ELRA, 2018. Available: http://www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf [last viewed 14.04.2024].

European Commission. "Licences for Europe" stakeholder dialogue. 22 December 2017. Available: https://digital-strategy.ec.europa.eu/en/library/licences-europe-stakeholder-dialogue [last viewed 14.04.2024].

Feng, Z. Past and Present of Natural Language Processing. In: Formal Analysis for Natural Language Processing: A Handbook. Singapore: Springer, 2023. Available: https://doi.org/10.1007/978-981-16-5172-4_1 [last viewed 14.04.2024].

Foster, D. Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play (Japanese Version). O'Reilly Media Incorporated, 2019.

Goldberg, Y. Features for Textual Data. In: Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies. Springer, Cham., 2017. Available: https://doi.org/10.1007/978-3-031-02165-7_6 [last viewed 14.04.2024].

Han, J., Pei, J., & Tong, H. Data mining: Concepts and Techniques. 4th edition. Morgan Kaufmann: 2022.

Heikkila, M. This new data poisoning tool lets artists fight back against generative AI. 2023. Available: https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai [last viewed 14.04.2024].

Hugenholtz, P. B., Quintais, J. P. Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output? IIC 52, 2021. Available: https://doi.org/10.1007/s40319-021-01115-0 [last viewed 14.04.2024].

Ilin, I. Legal Regime of the Language Resources in the Context of the European Language Technology Development. In: Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2019. Lecture Notes in Computer Science. *Vetulani, Z., Paroubek, P., Kubis, M.* (eds). Springer, Cham, Vol. 13212, 2022. Available: https://doi.org/10.1007/978-3-031-05328-3_24 [last viewed 14.04.2024].

Ilin, I., & Kelli, A. The use of human voice and speech in language technologies: the EU and Russian intellectual property law perspectives. Juridica International, Vol. 28, 2019, pp. 17–27.

Imamguluyev, R. The Rise of GPT-3: Implications for Natural Language Processing and Beyond. International Journal of Research Publication and Reviews, 4(3), 2023, pp. 4893–4903. Available: https://doi.org/10.55248/gengpi.2023.4.33987 [last viewed 14.04.2024].

Kamocki, P., Bond, T., Lindén, K., Margoni, T., Kelli, A., Puksas, A. Mind the Ownership Gap? Copyright in AI-generated Language Data, Linköping University Electronic Press (forthcoming), 2024.

Kamocki, P., Linden, K., Puksas, A., Kelli, A. EU Data Governance Act: Outlining a Potential Role for CLARIN. In: CLARIN Annual Conference 2022, *Erjavec, T., Eskevich, M.* (eds). Linköping Electronic Conference Proceedings, 2023. Available: https://doi.org/10.3384/ecp198006 [last viewed 14.04.2024].

Kamocki, P., Hannesschläger, V., Hoorn, E., Kelli, A., Kupietz, M., Lindén, K., Puksas, A. Legal Issues Related to the Use of Twitter Data in Language Research, 2022. In: Selected Papers from the CLARIN Annual Conference 2021. Linköping Electronic Conference Proceedings, *Monachini, M., Eskevichm, M.* (eds). Available: https://doi.org/10.3384/ecp1897 [last viewed 14.04.2024].

Kelli, A., Lepik, G. Originality as a Key Concept of the Estonian Copyright Law. In: Handbook on Originality in Copyright, *Gupta, I.* (eds). Singapore: Springer, 2023. Available: https://doi.org/10.1007/978-981-19-1144-6_10-1 [last viewed 14.04.2024].

Kelli, A., Mets, T., Jonsson, L., Pisuke, H., Adamsoo, R. The changing approach in Academia-Industry collaboration: From profit orientation to innovation support. Trames Journal of the Humanities and Social Sciences, 17(3), 2013, pp. 215–241. Available: http://doi.org/10.3176/tr.2013.3.02 [last viewed 14.04.2024].

Kelli, A., Tavast, A., Lindén, K. Building a Chatbot: Challenges under Copyright and Data Protection Law. In: Contracting and Contract Law in the Age of Artificial Intelligence, *Ebers, M., Poncibò, C., Zou, M.* (eds). Bloomsbury Publishing, 2022.

Kelli, A., Tavast, A., Lindén, K., Vider, K., Birštonas, R., Labropoulou, P., Kull, I., Tavits, G., Värv, A., Stranák, P., Hajic, J. The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies. In: Selected Papers from the CLARIN Annual Conference 2019, Linköping University Electronic Press., 2020. *Simov, K., Eskevich, M.* (eds). Available: http://doi.org/10.3384/ecp2020172008 [last viewed 14.04.2024].

King, M. R., & ChatGPT. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. Cellular and molecular bioengineering, 16(1), 2023, pp. 1–2. Available: https://doi.org/10.1007/s12195-022-00754-8 [last viewed 14.04.2024].

Lemley, M. A. How Generative AI Turns Copyright Upside Down. 2023, pp. 12–13. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4517702 [last viewed 14.04.2024].

Margoni, T. Saving research: Lawful access to unlawful sources under Art. 3 CDSM Directive? Kluwer Copyright Blog, 2023. Available: https://copyrightblog.kluweriplaw.com/2023/12/22/saving-research-lawful-access-to-unlawful-sources-under-art-3-cdsm-directive/ [last viewed 14.04.2024].

Mayer, C. W., Ludwig, S., & Brandt, S. Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models. Journal of Research on Technology in Education, 55(1), 2023, pp. 125–141. Available: https://doi.org/10.1080/15391523.2022.2142872 [last viewed 14.04.2024].

Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H. Reframing Instructional Prompts to GPTk's Language. 2021. Available: https://doi.org/10.48550/arXiv.2109.07830 [last viewed 14.04.2024].

Opinion of Advocate General Szpunar delivered on 25 October 2018, case No. C-469/17, Funke Medien NRW GmbH v. Bundesrepublik Deutschland, para. 1. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62017CC0469&qid=1669110125285 [last viewed 14.04.2024].

Pagallo, U., & Sciolla, J. C. Anatomy of web data scraping: ethics, standards, and the troubles of the law. European Journal of Privacy Law & Technologies, No. 2, 2023. Available: https://dx.doi.org/10.2139/ssrn.4707651[last viewed 14.04.2024].

Senftleben, M. Generative AI and Author Remuneration. International Review of Intellectual Property and Competition Law, 54, 2023, pp. 1535–1560. Available: https://doi.org/10.1007/s40319-023-01399-4 [last viewed 14.04.2024].

Truyens, M., Van Eecke, P. Legal aspects of text mining. Computer Law & Security Review, 30(2), 2014, pp. 153–170. Available: https://doi.org/10.1016/j.clsr.2014.01.009 [last viewed 14.04.2024].

Vincent, J. Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement. 6 February 2023. Available: https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion [last viewed 14.04.2024].

Weizenbaum, J. ELIZA – a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 1966, pp. 36–45. Available: https://doi.org/10.1145/365153.365168 [last viewed 14.04.2024].

Zhou, M., Duan, N., Liu, S., & Shum, H. Y. Progress in neural NLP: modeling, learning, and reasoning. Engineering, 6(3), 2020, pp. 275–290. Available: https://doi.org/10.1016/j.eng.2019.12.014 [last viewed 14.04.2024].

Zong, C., Xia, R., & Zhang, J. Text data mining, Vol. 711, Singapore: Springer, 2021. Available: https://doi.org/10.1007/978-981-16-0100-2 [last viewed 14.04.2024].

Normative acts

Agreement on Trade-Related Aspects of Intellectual Property Rights (as amended on 23 January 2017). Available: https://www.wto.org/english/docs_e/legal_e/31bis_trips_01_e.htm [last viewed 14.04.2024].

Berne Convention for the Protection of Literary and Artistic Works, signed at Berne on 9 September 1886 (Berne Convention). Available: https://www.wipo.int/wipolex/en/text/283698 [last viewed 14.04.2024].

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, 17.5.2019, pp. 92–125. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L0790&qid=1706810141810 [last viewed 14.04.2024].

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, 22.6.2001, pp. 10–19. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32001L0029&qid=1706810353762 [last viewed 31.01.2024].

Directive (EU) 2019/770 of the European Parliament and of the Council on certain aspects concerning contracts for the supply of digital content and digital services (Digital Content Directive) [2019] OJ L136/1. Available: https://eur-lex.europa.eu/eli/dir/2019/770/oj [last viewed 14.04.2024].

Estonian Copyright Act. Passed 11.11.1992. English translation available: https://www.riigiteataja.ee/en/eli/527122022006/consolide [last viewed 14.04.2024].

European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html [last viewed 14.04.2024].

Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). OJ L 152, 3.6.2022, pp. 1–44. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868&qid=1712322025563 [last viewed 14.04.2024].

WIPO Copyright Treaty. Adopted in Geneva on 20 December 1996. Available: https://www.wipo.int/wipolex/en/text/295166#P86_11560 [last viewed 14.04.2024].

Case law

Class Action Complaint and Demand for Jury Trial in case No. 3:22-cv-06823, Doe et al. v. GitHub, Inc. et al., United States District Court for the Northern District of California. Available: https://githubcopilotlitigation.com/pdf/06823/1-0-github_complaint.pdf [last viewed 14.04.2024].

Class Action Complaint and Demand for Jury Trial in case No. 1:24-cv-00084, Nicholas Gage v. Microsoft, OpenAI, United States District Court for the Southern District of New York. Available: https://fingfx.thomsonreuters.com/gfx/legaldocs/klvydkdklpg/OPENAI%20COPYRIGHT%20LAWSUIT%20basbanescomplaint.pdf [last viewed 14.04.2024].

Complaint and Demand for Jury Trial in case No. 1:23-cv-11195, the New York Times company v. Microsoft, OpenAI, United States District Court for the Southern District of New York. Available: https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf [last viewed 14.04.2024].

Judgment of the Court of Justice of the European Union of 11 June 2020, case No. C-833/18, SI, Brompton Bicycle Ltd v. Chedech/Get2Get. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=227305&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=1727690 [last viewed 14.04.2024].

Judgment of the Court of Justice of the European Union of 12 September 2019, case No. C-683/17, Cofemel – Sociedade de Vestuário SA v. G-Star Raw CV, para. 29. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=217668&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=9984025 [last viewed 14.04.2024].

Judgment of the Court of Justice of the European Union of 29 July 2019, case No. C-469/17, Funke Medien NRW GmbH v. Bundesrepublik Deutschland, para. 25. Available: https://curia.europa.eu/

juris/document/document.jsf?text=&docid=216545&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=10024671 [last viewed 14.04.2024].

Judgment of the Court of Justice of the European Union of 29 July 2019, case No. C-476/17, Pelham GmbH, Moses Pelham, Martin Haas v. Ralf Hütter, Florian Schneider-Esleben, para. 71. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=216552&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=547686 [last viewed 14.04.2024].

Judgment of the Court of Justice of the European Union of 16 November 2016, case No. C-301/15, Marc Soulier, Sara Doke v. Premier ministre, Ministre de la Culture et de la Communication, para. 34. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=185423&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=164343 [last viewed 14.04.2024].

Judgment of the Court of Justice of the European Union of 3 September 2014, case No. C-201/13, Johan Deckmyn, et al. v. Helena Vandersteen et al., para. 26. Available: https://curia.europa.eu/juris/document/document.jsf?mode=DOC&pageIndex=0&docid=157281&part=1&doclang=EN&text=&dir=&occ=first&cid=689582 [last viewed 14.04.2024].

Judgment of the Court of Justice of the European Union of 1 March 2012, case No. C-604/10, Football Dataco Ltd, et al. v. Yahoo! UK Ltd, et al. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=119904&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=2084171 [last viewed 14.04.2024].

Judgment of the Court of Justice of the European Union of 16 July 2009, case No. C-5/08, Infopaq International A/S v. Danske Dagblades Forening, para. 48, 51. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=72482&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=1892287 [last viewed 14.04.2024].

Other sources

AI Prompt Marketplace. Terms of service. Available: https://promptrr.io/terms-of-service/ [last viewed 14.04.2024].

Notice of inquiry and request for comments. U.S. Copyright Office, Library of Congress. Federal Register, Vol. 88, No. 167, Wednesday, August 30, 2023. Available: https://www.govinfo.gov/content/pkg/FR-2023-08-30/pdf/2023-18624.pdf [last viewed 14.04.2024].

Notion AI Supplementary Terms. Available: https://www.notion.so/notion/Notion-AI-Supplementary-Terms-fa9034c8b5a04818a6baf3eac2adddbb [last viewed 14.04.2024].

OpenAI. Terms of use. Effective: January 31, 2024. Available: https://openai.com/policies/terms-of-use [last viewed 14.04.2024].

Prompt Submission Guidelines, PromptBase. Available: https://promptbase.com/prompt-guidelines [last viewed 14.04.2024].

PromptBase. Terms of Service. Available: https://promptbase.com/tandcs [last viewed 14.04.2024].

U.S. Copyright Office, Letter 05.09.2023, Re: Second Request for Reconsideration for Refusal to Register Théâtre D'opéra Spatial (SR # 1-11743923581). Available: https://www.copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf [last viewed 14.04.2024].

U.S. Copyright Office, Letter 21.02.2023, Re: Zarya of the Dawn (Registration #VAu001480196). Available: https://www.copyright.gov/docs/zarya-of-the-dawn.pdf [last viewed 14.04.2024].