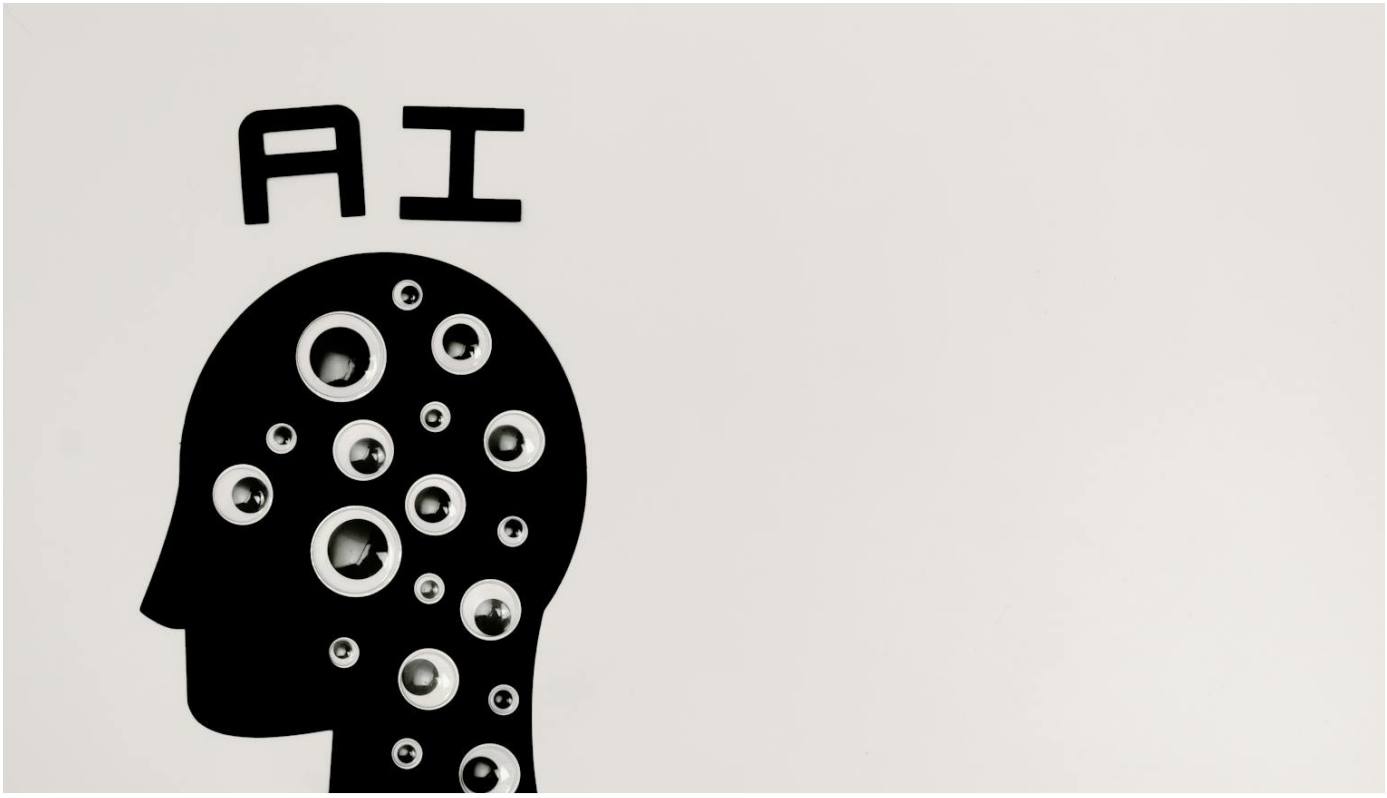

How should copyright-related concerns be tackled when “feeding” generative AI models?

9.9.2025

Rapidly emerging generative AI (GenAI) models, such as ChatGPT and Midjourney, have quickly become mainstream due to their impressive ability to generate different types of data promptly and economically. However, to function effectively, these systems must be trained on vast amounts of data that are often subject to copyright. Thus, the growth of copyright-related concerns in GenAI comes as no surprise.

AI developers employ text and data mining (TDM) to train their GenAI systems. This

technology enables the analysis of extensive volumes of digital data, so-called Big Data, stored worldwide. Various techniques (e.g., classification, clustering, keyword extraction, summarisation, etc.) can be used to extract patterns, insights, correlations, or other valuable information that enable AI systems to generate relevant outputs. Training datasets used to “feed” such systems usually consist of a wide range of creative works (e.g., images, music, books, etc.) that may be subject to copyright. Therefore, when a computer makes copies of such data in the course of TDM, it may infringe copyright holders’ exclusive right to reproduction (Art. 2 of the InfoSoc Directive).

To prevent copyright infringement, providers of GenAI models may obtain authorisation from rightholders to mine protected works through contractual arrangements, or they may choose to train their models on public domain works or data made available under open-source licenses. However, neither of these scenarios is ideal. Obtaining licenses for vast amounts of digital data may require significant financial resources and time, while restricting training data may reduce the quality of AI models. Against this background, AI developers may benefit from a specific TDM exception provided under Art. 4 of the Directive on Copyright in the Digital Single Market (CDSM Directive). The provision was adopted alongside another TDM exception introduced under Art. 3 of the CDSM Directive, which applies only to research organisations and cultural heritage institutions conducting TDM for scientific research. The exception of Art. 4 is much broader, designed to compensate for the narrow scope of Art. 3, as it permits all types of users with lawful access (both commercial and non-commercial) to carry out TDM for any purpose.

However, several aspects may dilute the effectiveness of this provision and hinder its practical application. Users, such as providers of GenAI, can benefit from the exception only if rightholders have not reserved the use of their works for TDM. Copyright holders may opt out of the exception by employing machine-readable means, including metadata and the terms and conditions of a website or service, in cases where their works are publicly available online. In other circumstances, they may reserve their rights through other means, such as contractual agreements or unilateral declarations (Recital 18 of the CDSM Directive). In this regard, users could face additional costs, as they would be required not only to pay for lawful access but also to pay for the right to mine (analyse) lawfully obtained data. Such a scenario could undermine the innovative potential of small IT firms with limited resources, while further strengthening the position of large AI actors in the EU.

Moreover, there are currently no generally accepted protocols or technical standards that could provide more clarity on the operation of an “opt-out” mechanism in practice.

Rightholders may, for instance, employ the Robots Exclusion Protocol (the so-called *robots.txt file*) to prevent AI crawlers from accessing and using their works for training. However, the protocol is limited in size, functions only as a voluntary measure informing web bots whether content can be indexed or scraped, and can easily be ignored by AI crawlers. Nowadays, many AI developers offer alternative “opt-out” solutions that enable rightholders to control the use of their works for AI training, such as the *spawningai.txt* file and the *HaveIBeenTrained* website. The latter allows authors to search for their works in training datasets and opt out. However, these solutions are model-specific, and it remains unclear whether they are compatible with the CDSM Directive’s reservation of rights.

The rightholder-oriented approach of the “commercial” TDM exception is further reinforced by Art. 7 of the CDSM Directive, which allows the exception to be overridden by contract. Granting authors and other rightholders excessive control over the use of digital data for TDM could undermine the right to TDM. The purpose of AI training is not to reproduce (create exact copies of) lawfully obtained materials in order to substitute authors’ works, but to generate new outputs by processing and analysing the ideas and facts embedded in them. In this sense, it may be reasonable to amend the “commercial” TDM exception by excluding the reservation right. This would broaden the scope of the exception and facilitate wider access to data that is crucial for the development and operation of cutting-edge AI systems.

About the Author

Maryna Manteghi is a doctoral researcher at the Faculty of Law of the University of Turku. Her research focuses on the interaction of AI technologies and copyright law. UTU webpage https://www.utu.fi/en/people/maryna-manteghi

#copyright law #generative AI #Text and data mining

Insights

The mission of the University of Turku's faculty of law online publication is to promote the visibility of the research conducted within the faculty. The aim is to highlight the faculty's activities through various current themes. Insights is created in collaboration with the University of Turku's communications team.

Contact
insights@utu.fi

Accessibility Statement

Privacy Notice

ISSN 2984-5246