Copyright Clearance Center

True or False?
Addressing Common Assumptions About Copyright and AI

Copyrighted materials are the fuel for artificial intelligence (AI) systems, but misunderstandings persist about how copyright applies to the use of content as training material for AI models. These misunderstandings extend to content used in end user applications of AI, such as the summarization of collections of articles, interrogation of documents for insights, automation of literature screening, and creation of visualizations of content sets, among others.

Let's explore some common assumptions about copyright and AI and address several widely held misunderstandings on the subject.

1 True or False? No copies are made during the process of training generative AI systems. The "machine" is only reading the content and learning from it, just like a human would.

False. There are a number of differences in the way humans and machines learn. Humans learn through a combination of direct experience, observation, imitation, and abstract thinking, while large language models (LLMs) "learn" from the processing of vast amounts of training data, which is copied and stored.

The process of training LLMs involves feeding the models enormous quantities of data, which can consist of text, images, and audio and video files. The use of copyrighted content is paramount because it contains high-quality, diverse materials covering a wide range of topics, styles, and complexities. These materials allow AI providers to create more versatile and robust models that in turn produce more reliable outputs.

LLMs retain expressions of the original works on which they have been trained. Generally, LLMs make and store local copies of their training materials to accelerate the learning process and provide access to the original dataset. The content is then "tokenized" (sentences are broken down to words and phrases, and words are broken down into characters) to ensure that the text data can be converted into a format that the LLMs can process, while preserving the semantic structure and meaning of the original work. LLMs recognize the context of words based on the expression of those words in the original work, and these systems are able to recall significant portions, and in some cases, entire works, verbatim. The copying and retention of these works in AI systems and their reproduction in outputs implicates copyright, making appropriate licensing essential for copyright compliance.

2 True or False? Using copyrighted content for training LLM systems is fair use.

It depends. At present, there are numerous copyright infringement cases brought by authors, publishers, and other rightsholders against AI companies regarding the use of their content in the training of LLMs without consent. These cases are still being litigated. Some AI companies defend their use of copyrighted materials as "fair use." In the United States, the fair use doctrine promotes the freedom of expression by allowing the unlicensed use of a copyrighted work under *certain* circumstances (e.g., criticism, commentary, news reporting, teaching, scholarship, or research).

Determining if a particular use qualifies as fair use is a highly fact-specific inquiry; each case is different and must be evaluated individually. There is no blanket rule or exception under fair use that uniformly applies to all AI-related activities. Securing appropriate licenses is the best way to gain permission for uses of copyrighted content, including for certain activities related to AI.

3 True or False? I can avoid all copyright infringement issues by only using open access (OA) content with AI systems.

False. When using OA content, it's important to understand the type of OA license under which the content is made available. A Creative Commons license, for example, is a type of OA license that itself has six main types, each granting a different set of permissions for use under a specific set of conditions, including attribution. While some of those licenses authorize use for commercial purposes, several specify that use is limited only for non-commercial purposes. In addition, some rightsholders use their own forms of OA licenses that have different terms from the Creative Commons licenses.

Although some OA licenses are more permissive, it is crucial to understand and adhere to the specific OA license terms.

4 True or False? Copyright stands in the way of innovation.

False. Copyright law in the United States was created for the purpose of promoting science and the useful arts. The U.S. Constitution recognizes in the Copyright Clause (Article 1, Section 8, Clause 8) that authors have the right to benefit from their creations for a limited time. Copyright fosters innovation by providing creators with financial incentives, legal protections, and recognition for their creative contributions. Copyright has long supported advancements in technology, from the invention of the photocopier through the development of the internet. Licensing, both direct and collective, provides a convenient way to respect copyright and support innovation. Appropriate remuneration for use of copyrighted material encourages investment in research and development, promotion of technological advancements, and support for diverse forms of creative expression. Copyright continues to play a vital role in driving innovation and economic growth in society.

Drive Business Forward with the Annual Copyright License

A copyright compliance strategy that informs and meets the needs of employees across the enterprise sets an organization up for higher efficiency, improved collaboration, and a minimized risk of copyright infringement, ultimately helping to fuel innovation and new discoveries. CCC's Annual Copyright License complements an organization's publisher agreements and subscriptions to enable teams to collaborate more easily using content from a wide range of sources, authorizes the internal-only use of lawfully acquired content with artificial intelligence (AI) systems, simplifies copyright compliance, drives innovation, and offers a library of resources to educate employees about the importance of copyright.

Click here to contact us about content management and licensing solutions for your organization.

Beth Johnson

Beth Johnson is Corporate Solutions Director at CCC. She is responsible for developing go-to-market strategies, conducting research, and developing positioning and messaging for the corporate copyright licenses. Beth's background is in medical publishing, managing product development from concept to maturity, across technologies and media in both emerging and established global markets. Before joining CCC she served in leadership roles at Greylock Press, SAGE Publications, The Goodwin Group International, and the Massachusetts Medical Society.

About CCC

A pioneer in voluntary collective licensing, CCC advances copyright, accelerates knowledge, and powers innovation. With expertise in copyright, data quality, data analytics, and FAIR data implementations, CCC and its subsidiary RightsDirect collaborate with stakeholders on innovative solutions to harness the power of data and AI.

Copyright Clearance Center

Learn more

For more information, please visit our AI, Copyright & Licensing insight page.

copyright.com/resource-library/communities/ai-copyright-licensing/