Study Guidelines

by Ralph NACK, Reporter General
Guillaume HENRY and Johanna FLYTHSTRÖM, Deputy Reporter General
Rafael ATAB, Klaudia BLACH-MORYSINSKA, Mamta Rani JHA and Yanfeng XIONG,
Assistants to the Reporter General

2025 - Study Question

Copyright and Artificial Intelligence

Introduction

1) Since AIPPI's pioneering Resolution "Copyright in Artificially-Generated Works", adopted at the 2019 London Congress, artificial intelligence ("AI") has continued to evolve rapidly, particularly with the development of Generative AI (“GENAI”), such as ChatGPT launched in 2022, Midjourney, or Stable Diffusion.

2) These new AI systems are developed through the large-scale collection of data available on the Internet (also known as web scraping), which is subsequently usually organized into structured datasets, for the purpose of training the AI System. However, some of the data scraped from the internet is protected by copyright. The issue therefore arises as to whether or not the scraping of copyrighted works from the Internet, and their subsequent use to train an AI System, is legal or not, *i.e.* whether such acts require prior consent from the copyright holders and whether the content generated by an AI system is infringing.

Why AIPPI considers this an important area of study

3) The use of copyrighted works to train AI Systems is currently one of the most hotly debated issues in the field of intellectual property.

4) The legal framework for AI, both at the national and international levels, is a work in progress. Numerous legislative initiatives are underway, and number of case law decisions are eagerly awaited in the coming months.

5) In this field, harmonisation is essential, as AI Systems are inherently global in their deployment and use.

6) The stakes are high, as the decision on whether copyrighted works can be used as training data could slow down the development of AI Systems or, conversely, cause considerable harm to authors and ultimately dry up human creation. The aim is therefore to find a balance that allows authors to make a living from their work and continue to create, while not hindering the development of AI.

Definitions

7) In the context of this study, the following terms have the following definitions:

a. The term "**Copyright**" means the rights associated with copyright as set forth in the Berne Convention AND all other copyright-type rights, *e.g.* "related rights", "neighbouring rights", etc. Other rights such as image right, privacy right, etc. are outside the scope of this Study Question.

b. The term "**Economic Rights**" means the exclusive rights of Copyright, *e.g.* the right of reproduction, representation, adaptation, etc.

c. The term "**Moral Rights**" means the rights of Copyright apart from Economic Rights, *e.g.* the right to object to distortion of the work, right to authorship, etc.

d. The term "**Exception(s)**" means any legal or jurisprudential exception or limitation to the Economic Rights and/or Moral Rights. An Exception can be subject or not to indemnification / compensation to Copyright holders.

e. the term "**AI System**" means a machine-based system (AI, GENAI, etc.) that:
   - ▪ is designed to operate with varying levels of autonomy and may exhibit adaptiveness after deployment; and
   - ▪ from the input (training data) it receives, generates outputs such as new contents, etc.

f. The term "**Copyrighted Data**" means one or more data (music, images, videos, texts, etc.), protected under Copyright law.

g. The term "**Use of Copyrighted Data to Train AI System**" means Copyrighted Data that are:
   - ▪ scraped off the Internet (or by any other means); and
   - ▪ possibly structured and incorporated into a dataset; and
   - ▪ used to train an AI System.

   These operations usually result in acts of reproduction of Copyrighted Data, but not absolutely systematically.

h. The Term "**Provider of an AI System**" means a natural or legal person that develops an AI System and places it on the market or puts the AI System into service.

Scope of this Study Question

8) The aim of this Study Question is to determine whether and under what conditions a Copyrighted Data can be used as training data to train an AI System and under which conditions a content generated by an AI System (output) and/or an AI System itself are considered to be infringing.

9) This Study Question also aims to determine the sanctions for using a Copyrighted Data as training data or in the content generated (output) without copyright holder consent.

Previous work of AIPPI

10) AI was a major topic of discussion during many Panel Sessions in the past years. For instance:

- A Panel Session on "Big Data" at the Sydney Congress in 2017;
- A dedicated, double-length Panel Session on "Artificial Intelligence – the Real IP Issues" at the Cancun Congress in 2018; and
- A Panel Session on "The Copyright Dilemma: Trained to Infringe?" at the Hangzhou Congress in 2024.

11) AIPPI adopted a landmark Resolution related to "Copyright in Artificially-Generated Works", at the London Congress in 2019. This 2019 Resolution determines if and under what conditions Copyright should be available for artificially-generated works. It was focused on protection of *outputs*. However, the 2019 Study Question did not address whether the use of copyrighted training data falls within the scope of Economic Rights, and whether such use constitute an infringement. This is the issue of *inputs*. The 2019 Study Question did not address either under what conditions the outputs can constitute a Copyright infringement.

12) Furthermore, in 2020, AIPPI adopted a Resolution on "IP Rights in Data". This Resolution addressed the issue of rights in data, in particular IP rights in structured and unstructured data under existing or possible new forms of protection. But this Resolution did not address the Use of Copyrighted Data to Train AI Systems.

Discussion

13) AI, and particularly GENAI, undergoes extensive training on large datasets to recognize patterns and relationships within the data. Consequently, the initial design phase of the AI System, prior to the training phase, involves massive data collection, often through Text and Data Mining (“TDM”), usually on the Internet. Data is usually downloaded, processed and stored into structured datasets. This is followed by the training phase, during which the AI System uses this input data to develop its capabilities.

14) Consequently, the performance of an AI System and the quality of the contents generated (output) are fundamentally dependent on the quality of the input data. While this is not an absolute principle, Copyrighted Data often tends to be of higher or more reliable quality than non-Copyrighted Data.

15) In the European Union, the legal framework governing the use of Copyrighted works to train an AI System has already been partially developed. Directive 2019/790 of 17 April 2019 on copyright and related rights in the Digital Single Market provides limitation and exceptions to Copyright to allow Text and Data Mining (TDM) for the

purposes of scientific research[1] and more generally for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of TDM[2] . In the latter case, an opt-out system has been introduced, allowing the author to reserve the application of this Exception.

Recital 105 of the Regulation EU 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (”AI Act") states that: *"General-purpose AI models, in particular large generative AI models, capable of generating text, images, and other content, present unique innovation opportunities but also challenges to artists, authors, and other creators and the way their creative content is created, distributed, used and consumed. The development and training of such models require access to vast amounts of text, images, videos and other data. Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights. Any use of copyright protected content requires the authorisation of the rightsholder concerned unless relevant copyright exceptions and limitations apply. <u>Directive (EU) 2019/790 introduced exceptions and limitations allowing reproductions and extractions of works or other subject matter, for the purpose of text and data mining, under certain conditions</u>. Under these rules, <u>rightsholders may choose to reserve their rights over their works or other subject matter to prevent text and data mining</u>, unless this is done for the purposes of scientific research. Where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightsholders if they want to carry out text and data mining over such works"*. Furthermore, Article 53.1 provides that: *"Providers of general-purpose AI models shall (...) (d) draw up and make publicly available a <u>sufficiently detailed summary about the content used for training of the general-purpose AI model</u>, according to a template provided by the AI Office".*

16) In the UK, Getty Images claims that Stability AI has copied more than 12 million photographs from Getty Images' collection, along with the associated captions and metadata, without permission from or compensation to Getty Images[3].

[1] *"Article 3. Text and data mining for the purposes of scientific research.*
1. Member States shall provide for an exception to the (copyright) (...) <u>for reproductions and extractions</u> made by <u>research organisations and cultural heritage institutions</u> in order to carry out, for the <u>purposes of scientific research, text and data mining of works</u> or other subject matter to which they have lawful access (...)".

[2] *"Article 4. Exception or limitation for text and data mining.*
1. Member States shall provide for an exception or limitation to (copyright) (...) for reproductions and extractions of lawfully accessible works and other subject matter <u>for the purposes of text and data mining.</u>
2. Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining.
3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph <u>has not been expressly reserved by their rightholders</u> in an appropriate manner, such as machine-readable means in the case of content made publicly available online."

[3] https://www.courtlistener.com/docket/66788385/getty-images-us-inc-v-stability-ai-inc/

17) In India, Asian News International (ANI) has filed Copyright infringement suit before the Delhi High Court against OpenAI Inc., alleging unauthorized use of original news content to train OpenAI 's LLM.

18) In the US, an initial order has been issued by the Californian court in proceedings between Midjourney, Stable Diffusion and DeviantArt and three artists, alleging infringement of their Copyright in the training and use of GENAI systems[4].

In 2023, the New York Times sued OpenAI and Microsoft for the illicit use of Times articles to train GENAI. Times claims that OpenAI is infringing on Copyright through the unlicensed and unauthorized use and reproduction of its works during the training of its models. This case could have a significant impact, particularly with respect to fair use. On November 7, 2024, New York court dismissed the lawsuit brought by two news sites (Raw Story and AlterNet) that claimed OpenAI has violated the Digital Millennium Copyright Act by removing author and Copyright information from their articles used as training data for ChatGPT, considering that the plaintiffs had not sufficiently alleged harm caused by the removal of author information from ChatGPT training sets.

The proposed "Generative AI Copyright Disclosure Act of 2024" of 9 April 2024 proposes to require companies to disclose training data for their GENAI systems, including those already on the market[5] : *"A person who creates a training dataset, or alters a training dataset (including by making an update to, refining, or retraining the dataset) in a significant manner, that is used in building a generative AI system shall submit to the Register a notice that contains- (A) a sufficiently detailed summary of any copyrighted works used - (i) in the training dataset (in the case that the person creates the dataset); or (ii) to alter the training dataset (in the case that the person alters the training data in a significant manner); and (B) the URL for such dataset (in the case of a training dataset that is publicly available on the internet at the time the notice is submitted)".*

19) The rapid emergence of AI has outpaced the development of a comprehensive legal framework, which remains incomplete even in the European Union, where implementing regulations are still being drafted. Various jurisdictions adopt differing strategies, reflecting the timeless dilemma that arises with the advent of new technologies: whether to legislate promptly or await judicial decisions.

The first option is to legislate quickly to establish a stable legal framework, in order to foster the development of AI systems, while ensuring compliance with Copyright law. This is the approach adopted by the European Union, where the legal framework seeks to balance the interests of AI developers with those of Copyright holders. It encourages stakeholders to negotiate agreements (collecting societies will probably play an important role in this legal framework). However, this option runs the risk (as

[4] https://www.courtlistener.com/docket/66732129/andersen-v-stability-ai-ltd/

[5] https://www.congress.gov/bill/118th-congress/house-bill/7913/text

always) of creating a legal framework that could quickly become obsolete as technology evolves and, if overly complex, could hinder AI development in Europe.

The second option is to rely on case law, applying and interpreting existing legal rules, to address the challenges posed by new technology. For the time being, this is the choice made by the majority of jurisdictions. However, this option entails (as always) the risk of having to wait a long time for the supreme courts to issue their rulings.

The proposed US Generative AI Copyright Disclosure Act of 2024 of 9 April 2024 is an intermediate solution, imposing a transparency obligation on AI systems, while leaving it to judges to decide whether AI systems comply with Copyright law.

20) Whichever option is chosen, it should be once again emphasised that the optimal development of AI systems in compliance with Copyright law can only be achieved with harmonised legal rules. That is the ambition of this Study Question: to propose a balanced legal framework for the use of training data protected by Copyright.

You are invited to submit a Report addressing the questions below. Please refer to the 'Protocol for the preparation of Reports'.

Questions

To answer all questions, please refer to the definition part above, especially for the terms "Copyrighted Data" and "Use of Copyrighted Data to Train AI System".

I. Current law and practice

Please answer all questions in Part I on the basis of your Group's current law and practice.

1) Does your current law / practice contain laws, rules, regulations or case law decisions specifically relating to the Use of Copyrighted Data to Train AI System?
   Please answer YES or NO. Please explain.

Does the Use of Copyrighted Data to Train AI System have to be authorised by the Copyright holder?

2) Does the Use of Copyrighted Data to Train AI System fall within the scope of the Economic Rights monopoly, *i.e.* does it require the authorisation of the Copyright holder unless covered by an Exception? Please answer YES or NO. Please explain, *e.g.* if the authorisation is necessary only if acts of reproduction, representation, etc. are carried out, or in any circumstances (even if no act of reproduction, representation, communication, etc. has been carried out).

3) Are there any Exceptions that authorise the Use of Copyrighted Data to Train AI System without Copyright holder consent?

   a. General Exceptions to Copyright, *i.e.* which are not special to the Use of Copyrighted Data to Train AI System (*e.g.* fair use, intermediate storage/temporary reproduction, etc.)? Please answer YES or NO. Please explain, *e.g.* the conditions of application of each Exception separately.

   b. Special Exceptions to the Use of Copyrighted Data to Train AI System (*e.g.* Text and Data Mining -TDM Exception, etc.)? Please answer YES or NO. Please explain *e.g.* the conditions of application of each Exception separately.

4) Do the Exceptions provide for financial compensation for Copyright holders (*e.g.* a royalty paid to a collecting society, etc.)? Please answer YES or NO. Please explain.

5) Can the author object to Use of Copyrighted Data to Train AI System on the basis of his/her Moral Rights (for example on the basis of the right to integrity, paternity, etc.)?

6) Does the Provider of an AI System have to make public the training data used to train the AI System (transparency obligation)? Please answer YES or NO. If YES, please explain the degree of detail required (all works, categories according to sources - websites, etc.) and whether the developer / operator must keep the list of training data or the training data itself for a certain period of time?

The consequences of unauthorised Use of Copyrighted Data to Train AI System

To answer questions 7 to 10, please consider that the Use of Copyrighted Data to Train AI System falls within the scope of the Economic Rights monopoly (no Exception can be invoked) but is made without the consent of the Copyright holder.

Infringing items

7) Can the contents created (output) by an AI System be qualified as Copyright infringement in the following cases?

   a. The output contains characteristic elements of one or more Copyrighted Data to Train AI System? Please answer YES or NO. Please explain if needed.

   b. The output is in the same style as one or more Copyrighted Data used to train AI System? Please answer YES or NO. Please explain if needed.

   c. The output is in all cases infringing if it has been trained with one or more infringed Copyrighted Data. Please answer YES or NO. Please explain if needed.

   d. In other cases? Please answer YES or NO. Please explain if needed.

8) Can the AI System itself be considered to infringe Copyright? Please answer YES or NO. If YES, please explain, *e.g.* under what conditions.

Infringer

9) Who is liable in case of Copyright infringement?

   a. The Provider of an AI System? Please answer YES or NO. Please explain if needed.

   b. The user who exploits commercially the AI System? Please answer YES or NO. Please explain if needed.

   c. The final user? Even if acting in good faith or unaware of the infringement? Please answer YES or NO. Please explain if needed.

   d. Any other person? Please answer YES or NO. Please explain if needed.

Sanctions

10) What sanctions can be imposed if it is found that Copyright has been infringed in order to train the AI System (because Copyrighted Data has been used with no authorisation)?

   a. Injunction, destruction, etc. of Copyrighted Data used to Train AI System still present in the data set or in the AI System? Please answer YES or NO. Please explain if needed.

b. Injunction, destruction, etc. of the AI System itself deemed to be infringing, because it has been trained with infringed Copyrighted Data? Please answer YES or NO. Please explain.

c. Injunction, destruction, recall from commercial channels, etc. of outputs deemed to be infringing? Please answer YES or NO. Please explain if needed.

d. The award of damages, including punitive damages? Please answer YES or NO. Please explain if needed.?

e. Confiscation of all or part of the profits generated by the operation of the AI System? Please answer YES or NO. Please explain if needed.

f. Any other sanctions? Please answer YES or NO. Please explain if needed.

In which other situations can outputs be qualified as Copyright infringement?

11) Please explain, if and under what conditions outputs generated by an AI System are qualified as infringement of a Copyrighted work, *e.g.* because the output contains characteristic elements of the Copyrighted work, in the following cases:

a. In case the Use of the Copyrighted Data (work) to Train the AI System is covered by an Exception. Please answer YES or NO. Please explain if needed.

b. In case the Copyrighted work has NOT been used to train the AI System. Please answer YES or NO. Please explain if needed.

c. In case the Use of the Copyright Data to Train AI System has been authorised, is the content generated (output) always licit, even if some for instance outputs contain characteristic elements of the Copyrighted work? Please answer YES or NO. Please explain if needed.

12) Who is liable in case outputs infringes Copyright?

a. The Provider of an AI System? Please answer YES or NO. Please explain if needed.

b. The user who exploits commercially the AI System? Please answer YES or NO. Please explain if needed.

c. The final user? Even if acting in good faith or unaware of the infringement? Please answer YES or NO. Please explain if needed.

d. Any other person? Please answer YES or NO. Please explain if needed.

II. Policy considerations and proposals for improvements of your Group's current law

13) Could any of the following aspects of your Group's current law or practice relating to the Use of Copyrighted Data to Train AI System be improved? If YES, please explain.

a. Use of Copyrighted Data to Train AI System require prior authorisation of the Copyright holder? Please answer YES or NO. Please explain if needed.

b. Exceptions authorising the Use of Copyrighted Training Data without Copyright holder consent. Please answer YES or NO. Please explain if needed.

c. Consequences of illicit Use of Copyrighted Data to Train AI System.

i. What can be qualified as infringing products, *e.g.* outputs and/or AI System itself? Please answer YES or NO. Please explain if needed

ii. The sanctions that should be available in case an AI System has been recognised to have been trained with infringed Copyrighted Data without authorisation? Please answer YES or NO. Please explain if needed.

14) Are there any other policy considerations and/or proposals for improvement to your Group's current law falling within the scope of this Study Question? Please answer YES or NO. If YES, please explain.

III. Proposals for harmonisation

Please consult with relevant in-house / industry members of your Group in responding to Part III.

15) In your opinion, should Use of Copyrighted Data to Train AI System be harmonised? Please answer YES or NO. For what reasons?

If YES, please respond to the following questions without regard to your Group's current law or practice.

Even if NO, please address the following questions to the extent your Group considers your Group's current law or practice could be improved.

Should the Use of Copyrighted Data to Train AI System be authorised by the Copyright holder?

16) Should the Use of Copyrighted Data to Train AI System fall within the scope of the Economic Rights monopoly, *i.e.* should it require the authorisation of the Copyright holder as a matter of principle, unless covered by an Exception?
Please answer YES or NO. Please explain, *e.g.* if the authorisation should be necessary only if acts of reproduction, representation, etc. are carried out, or in any circumstances.

17) Should there be Exceptions that allow the Use of Copyrighted Data to Train AI System without Copyright holder consent:

a. General Exceptions to Copyright, *i.e.* which are not special to the Use of Copyrighted Data to Train AI System (*e.g.* fair use, intermediate storage / temporary reproduction, etc.)? Please answer YES or NO. Please explain, *e.g.* the conditions of application of each Exception separately.

b. Special Exceptions to the Use of Copyrighted Data to Train AI System (*e.g.* TDM Exception, etc.)? Please answer YES or NO. Please explain *e.g.* the conditions of application of each Exception separately.

18) Should the Exceptions provide for financial compensation for Copyright holders (*e.g.* a royalty paid to a collecting society, etc.)? Please answer YES or NO. Please explain.

19) Should the author be able to object to the Use of Copyrighted Data to Train AI System on the basis of his / her Moral Right (*e.g.* on the basis of the right to integrity, paternity, etc.)? Please answer YES or NO. Please explain if needed.

20) Should the Provider of an AI System be obliged to make public the training data used to train the AI System (transparency obligation)? Please answer YES or NO.
If YES, please explain, *e.g.* the degree of detail that should be required (all works, categories according to sources - websites, etc.) and whether the company should keep the list of training data or the training data itself for a certain period of time?

The consequences of unauthorised Use of Copyrighted Data to Train AI System

To answer questions 21 to 24, please consider that the Use of Copyrighted Data to Train AI System falls within the scope of the Economic Rights monopoly (no Exception can be invoked) but is made without the consent of the Copyright holder.

Infringing items

21) Should the contents created (output) by the AI System be qualified as Copyright infringement in the following cases?

a. The output contains characteristic elements of one or more Copyrighted Data to Train AI System? Please answer YES or NO. Please explain if needed.

b. The output is in the same style as one or more Copyrighted Data used to Train AI System? Please answer YES or NO. Please explain if needed.

c. The output is in all cases infringing if it has been trained with one or more infringed Copyrighted Data. Please answer YES or NO. Please explain if needed.

d. In other cases? Please answer YES or NO. Please explain if needed.

22) Should the AI System itself be considered a Copyright infringement? Please answer YES or NO. If YES, please explain, *e.g.* under which conditions.

Infringer

23) Who should be liable in case of Copyright infringement by the outputs ?

a. The Provider of an AI System? Please answer YES or NO. Please explain if needed.

b. The user who exploits commercially the AI System? Please answer YES or NO. Please explain if needed.

c. The final user? Even if acting in good faith or unaware of the infringement? Please answer YES or NO. Please explain if needed.

d. Any other person? Please answer YES or NO. Please explain if needed.

Sanctions

24) What sanctions should be imposed if it is found that Copyright has been infringed in order to train the AI System, because Copyrighted Data has been used with no authorisation?

a. Injunction, destruction, etc. of Copyrighted Data used to train AI System still present in the dataset or in the AI System? Please answer YES or NO. Please explain if needed.

b. Injunction, destruction, etc. of the AI System itself deemed to be infringing, because it has been trained with infringed Copyrighted Data? Please answer YES or NO. Please explain.

c. Injunction, destruction, recall from commercial channels, etc. of outputs found to be infringing? Please answer YES or NO. Please explain if needed.

d. Award of damages, including punitive damages? Please answer YES or NO. Please explain if needed.

e. Confiscation of all or part of the profits generated by the operation of the AI System? Please answer YES or NO. Please explain if needed.

f. Any other sanctions? Please answer YES or NO. Please explain if needed.

In which other situations should outputs be qualified as Copyright infringement?

25) Please explain, in the following cases, if and under what conditions outputs generated by an AI System should be qualified as infringement of a Copyrighted work, *e.g.* because the output contains characteristic elements of the Copyrighted work, or is in the same style as the Copyrighted Work, etc.:

a. In case the use of the Copyrighted work to train the AI System is covered by an Exception. Please answer YES or NO. Please explain if needed.

b. In case the Copyrighted work has NOT been used to train the AI System. Please answer YES or NO. Please explain if needed.

c. In case the Use of the Copyright Data to Train AI System has been authorized, is the content generated (output) always licit, even if some for instance outputs contain characteristic elements of the Copyrighted work? Please answer YES or NO. Please explain if needed.

26) Who should be liable in case of Copyright infringement by the outputs?

a. The Provider of an AI System? Please answer YES or NO. Please explain if needed.

b. The user who exploits commercially the AI System? Please answer YES or NO. Please explain if needed.

c. The final user? Even if acting in good faith or unaware of the infringement? Please answer YES or NO. Please explain if needed.

d. Any other person? Please answer YES or NO. Please explain if needed.

27) Please comment on any additional issues concerning any aspect of the Use of Copyrighted Data to Train AI System you consider relevant to this Study Question.

28) Please indicate which industry sector views provided by in-house counsel are included in your Group's answers to Part III.