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 202

2)   2, Midjourney, or Stable Diffusion.

3)   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. 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. Under these rules, rightsholders may choose to reserve their rights over their works or other subject matter to prevent text and data mining, 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 sufficiently detailed summary about the content used for training of the general-purpose AI model, 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) (...) for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works 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 for the purposes of text and data mining.
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 has not been expressly reserved by their rightholders 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[1].

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[2] : *"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

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

---

[1] https://www.courtlistener.com/docket/66732129/andersen-v-stability-ai-ltd/
[2] https://www.congress.gov/bill/118th-congress/house-bill/7913/text

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.

    NO. Finland has implemented Articles 3 and 4 of Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market (the "Copyright DSM Directive") by enacting Section 13b in the Finnish Copyright Act (404/1961, as amended). It can be interpreted to be applicable for training AI but was not enacted specifically for the purpose.

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 <u>Economic</u> 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 <u>only</u> if acts of reproduction, representation, etc. are carried out, or <u>in any circumstances</u> (even if no act of reproduction, representation, communication, etc. has been carried out).

    YES, if it infringes the Economic Rights of authors or rightsholders.

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

    a. <u>General Exceptions to Copyright</u>, *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 <u>separately</u>.

    NO. There is an Exception concerning intermediate reproduction (Section 11 a of the Finnish Copyright Act) but as it must not have independent economic significance, it would be difficult to apply this Exception to training of AI which, in most cases, does have important economic significance.

b. <u>Special Exceptions</u> 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 <u>separately</u>.

YES. Finland has implemented Articles 3 and 4 of Copyright DSM Directive t by enacting Section 13b in the Finnish Copyright Act. Text and data mining ("TDM") Exception seems be to be applicable. It is not perfectly fit for purpose, but it has been widely recognised as a possibility. In addition, EU AI Act has a presumption that it could be at least one of the alternatives for allowing the training of large language models.

The TDM Exception in the Finnish Copyright Act states firstly that anyone who has lawful access to a work may make copies of it for TDM purposes, and store the copies solely for that purpose, unless the author has expressly and appropriately reserved this right. Secondly, it is stated that research organizations and cultural heritage institutions that have lawful access to a work may also make copies of it for TDM in scientific research and store them for scientific research purposes, including later verification of research results, provided that the copies of the work are only accessible to those entitled to them. With regard to TDM conducted by research organizations and cultural heritage institutions for scientific research purposes, there is no opt-out right for the author.

According to the preparatory works of the Finnish Copyright Act, the "text and data mining" is to be understood as defined in Art. 2 of the Copyright DSM Directive, i.e., "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". However, under national and international copyright law, copyright exceptions should be interpreted narrowly.

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

NO. There are no national rules for compensation for Copyright holders where use of a Copyright protected work is based on the TDM Exception. The TDM Exception in the Finnish Copyright Act, as well as the definition of text and data mining is materially the same as Art. 3 and 4 of the Copyright DSM Directive.

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

NO. The Finnish Copyright Act recognizes right to paternity and right to integrity.

Right to paternity – The name of the author must be stated in a manner required by "good practice" when copies of a work are made, or when the work is made available to the public. However, this right is not absolute and depends on industry practices (e.g., authors of computer programs are typically not credited).

Right to integrity – The work cannot be modified or made available to the public in a way that is prejudicial to the author's reputation.

The Exceptions set out in the Finnish Copyright Act do not limit the Moral Rights of authors. Moral Rights are unlikely to be enforceable against *training* of AI Systems as the training is not prejudicial to the author's reputation and as crediting authors is not customary in technical processes such as AI training.

However, publishers and authors are concerned about AI affecting the relationship between the authors and the work at a later phase. It cannot be excluded that an AI System is used or trained to produce content that does infringe the Moral Rights of the authors.

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?

   NO, NOT YET. According to Art. 53 of the EU AI Act, the AI Office will provide a template for making a summary of the training data used when training a general-purpose AI system. The national implementation of the EU AI Act is pending and there is uncertainty with respect to exact scope of disclosure obligations.

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.

      YES. Although there is no Finnish case law regarding infringing outputs yet, the infringement would be considered under basic copyright principles. According to the CJEU, parts of a work enjoy copyright protection, provided that they contain elements which are the expression of the intellectual creation of the author of the work (C-5/08 Infopaq, para. 39).

   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.

NO. Style itself is not protected under the Finnish Copyright Act, as copyright protects expression, not the idea or style itself. Such output could even in certain cases fall under the parody, caricature, and pastiche Exception (Section 23 a of the Finnish Copyright Act).

Further, Section 51 of the Finnish Copyright Act prohibits making works available to the public by using a name that may easily be confused with a previous work or its author, but it does not prohibit works that are similar with prior works.

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

    NO. The fact that an AI model was trained on infringing material does not automatically make all its outputs infringing. Copyright infringement requires that the output reproduces essential elements of the original work, as mentioned above (see question 7a). If the output is regarded as independent from the previous work, it does not constitute infringement (Section 4.2 of the Finnish Copyright Act).

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

    NO.

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

YES. If it is a reproduction of copyrighted work or a work made available to the public with or without alterations, and none of the Exceptions of Copyright law apply. For example, if the code of the AI System is or contains a reproduction of Copyright-protected code, the AI System itself can be considered to infringe Copyright.

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.

    YES, if the Provider of an AI System is responsible for the process of training and if an AI System is trained on Copyrighted Data without permission or if none of the Exceptions apply.

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

    YES, if the user uses infringing outputs in violation of copyright. If a work is used wilfully or out of negligence in an infringing manner, the infringer may be liable

to pay damages for other loss, such as mental suffering and other detriment, in addition to reasonable compensation (Section 57, Subsections 1 and 2 of the Finnish Copyright Act).

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

YES, if the user conducts infringing acts, i.e., makes temporary or permanent reproductions of Copyrighted Data or makes it available to the public with or without modifications. Lack of awareness does not automatically exempt a user from liability. However, other Exceptions may apply, e.g., if the final user only generates an infringing output for private use (Section 12 of the Finnish Copyright Act) or if the parody, caricature, or pastiche Exception applies (see Section 23 a of the Finnish Copyright Act).

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

YES. Any other person that causes a Copyright infringement can be liable, such as a distributor distributing infringing material.

Sanctions

10) What <u>sanctions</u> 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.

YES. According to Section 56 g of the Finnish Copyright Act, if a person infringes Copyright, the court may prohibit the infringer from proceeding with or repeating the infringing act.

The court can order the destruction of an infringing copy or equipment used to create such copy, such as Copyrighted Data used to Train AI System.

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.

YES. See response above to subsection a. The court could potentially deem the AI System itself to infringe Copyright, if it is or contains reproductions of copyrighted works. However, destruction or injunction regarding the use of

the AI System is not available merely because Copyrighted Data has been used to train the AI System, if the AI System itself is not infringing.

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

YES. See response above to subsection a.

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

The award of damages YES, punitive damages NO. According to Section 57 of the Finnish Copyright Act, Copyright infringer can be obliged to pay reasonable compensation to the author. If Copyright is infringed wilfully or negligently or by committing a Copyright offence or violation, the infringer can also be obliged, in addition to pay reasonable compensation, pay damages to the author for any other loss, including mental suffering and other detriment .Under Finnish law, the damages awarded always depend on the case-by-case evaluation, but cover the full economic compensation for the loss which is caused by the infringement, which in practice often is the normal license fee plus any other harm caused by the infringement.

The Finnish legal system does not recognize punitive damages.

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.

YES. The Finnish Copyright Act does not expressly address the possible confiscation of the profits generated by the operation of the AI System. According to section 58.2§ of the Finnish Copyright Act, if the copyright infringement constitutes a criminal office and the court orders confiscation of criminal assets (being the infringing work), then the copyright owner can, at the same time demand that the infringing work is destroyed. Confiscation of profits is a possibility in criminal proceedings but is seen as more of a sanction than a return of unjust gains to the state.

In civil cases profit generated by the operation of the AI System can potentially be payable as damages (loss of profit) to the Copyright holder. In addition, the Finnish legal system recognizes a principle of return of unjust enrichment which could also potentially be invoked to recover profits generated by the operation of the AI System to the Copyright holder.

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

YES. Copyright offence is punishable under Chapter 49, Section 1 of the Finnish Penal Code and may lead to a fine or imprisonment up to two years, and

Copyright violation is punishable under Section 56 a of the Finnish Copyright Act and may lead to a fine. The Finnish Penal Code also includes provisions on related offences, such as circumvention of technical protection.

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

11) Please explain, if and under what conditions <u>outputs</u> 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 <u>Exception</u>. Please answer YES or NO. Please explain if needed.

   YES. The general rules regarding Copyright infringement apply, and although it is not likely, the output can in theory qualify, e.g., as an unauthorized reproduction of the Copyrighted work. Relying on an Exception may allow the use of the Copyrighted work to Train the AI System, but unless otherwise specifically authorized by the Copyright holder, does not authorize the reproduction of the Copyrighted Data as an output of the trained AI System.

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

   YES. Even if the Copyrighted work in question has not been used to train the AI System, the output can be infringing if the user has deliberately prompted or otherwise used the AI System to create infringing output.

   However, in some cases it could be argued that the output qualifies as an independent creation and not as an unauthorized reproduction of Copyrighted work, if the author has not deliberately tried to copy an existing Copyrighted work.

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

   NO. The general rules regarding Copyright infringement apply, and the output can qualify, e.g., as an unauthorized reproduction of the Copyrighted Data.

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.

   YES. The Provider of an AI System can be held liable for Copyright infringement, if the AI System itself causes the infringement by reproducing a

Copyrighted work. This situation can arise, e.g., where only limited amount of Copyrighted Data has been used to train the AI System leading to the AI System reproducing the Copyrighted Data as an output.

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

YES. The user who exploits commercially the AI System can be held liable for Copyright infringement, if the user causes the Copyright infringement, e.g., by distributing output that qualifies as an unauthorized reproduction of Copyrighted work.

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

YES. The final user of output generated by the AI System can be held liable for Copyright infringement, e.g., if the final user provides a prompt that causes the unauthorized reproduction of a Copyrighted work. The consequences of the infringement depend on whether the final user has acted wilfully or negligently. Reasonable compensation may be payable to the Copyright holder for the use of the Copyrighted work irrespective of the final user's good faith or awareness of the infringement.

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

YES. Any other party that causes a Copyright infringement, such as a distributor distributing an infringing output generated by the AI System.

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.

YES. Authorisation of the Copyright holder is always required if none of the Exceptions and limitations apply. However, the rules related to the practical side of receiving the required authorisation could be improved. Currently, the definition of "*lawful use*", valid "*reservation*" and mechanisms used for obtaining authorisation are subject to interpretation. Some level of debate has been going on, e.g., related to the applicability of licensing mechanisms or to authorising use of Copyrighted Data to Train AI System.

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

YES. Under the Finnish Copyright Act, the text and data mining Exception allows the use of copyrighted works for text and data mining if the user has lawful access to the work and the right of reproduction has not been expressly reserved by the copyright holder in an appropriate manner.

Clarity is desirable. There is some unclarity on whether the text and data mining Exception in fact applies to Use of Copyrighted Data to Train AI System, as the literal definition of text and data mining only applies to actions aiming at producing analytical information. The current common practice of interpreting the text and data mining Exception allowing Training of AI Systems stems from the recital 105 of the AI Act, and the interpretation could be called to question. Consequently, there exists a demand for case law clarifying the definition of text and data mining, or alternatively, more precisely worded new legislation.

Furthermore, clarification on the acceptable means of reserving the copyright holder's rights fulfilling the "express" and "appropriate" requirements would be needed, as the DSM Directive refers to "machine readable means in the case of content made publicly available online" in Article 4 as an example, it is not clear what the format for executing this should be. Case law or development of common market practices in EU can solve the issue only within EU. International mechanisms for opt out would be preferable.

c. <u>Consequences</u> 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

   Yes. Reproductions of works, whether or not a part of AI Systems, can qualify as infringing. Output can qualify as infringing when it reproduces a work or a part thereof. However, it remains unclear if the use of AI System to produce infringing output can qualify as an infringement if it does reproduce a work in whole or in part and this results from use of AI System that is trained in accordance with applicable Exceptions or with author's consent that contains no agreement concerning output.

   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.

   YES. While current sanctions may well be sufficient, further clarity and specificity in applying the current law in AI System context would be needed in form of case law especially in relation to establishing an effective basis for claiming compensation in connection with the fault consisting of the lack of compliance with a duty of care under law: it has been publicly declared (in connection with the preparation of the now-suspended proposed AI liability directive, Article 4) that it can be challenging for claimants to establish a causal link between such non-compliance and the output

produced by the AI System or the failure of the AI System to produce an output that gave rise to the relevant damage.

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.

YES. Debate is ongoing in various organizations, but there are no clear outcomes yet.

In Finland, collective societies have begun to collect authorizations from rights holders to represent rights holders in cases related to AI training. In addition, some publishers have started to claim rights from the authors conserving all utilization of copyright-protected works related to AI Systems, in their agreements with authors. It seems as if different interest groups are reacting to the legal challenges caused by AI Systems and training thereof and are developing strategies on how to approach these challenges on a concrete level.

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.

YES.

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.

YES.

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.

      NO. General Exceptions are not necessary if special Exceptions apply

   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.

      YES.

      In absence of new legislation, the TDM Exceptions could provide an appropriate framework for regulating the use of copyrighted data to train AI systems, provided, however, that the scope of these Exceptions is clarified. Specifically, clarifying the scope to explicitly include the training of AI systems would help ensure consistent and transparent application of these Exceptions.

      An applicable Exception should define the retention period for reproductions and extractions to also cover the storage required for training AI systems.

      The possibility for rights holders to opt out could be an effective solution for copyright holders to retain control over their rights. However, there should be a defined, standardised, and easily accessible mechanism for rights holders to manage their preferences regarding the use of their works. This mechanism should also be easily verifiable by users who train AI Systems using the materials.

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.

   YES, if a license is needed. Depending on the mechanism and the purpose of use of the AI System, the possibility for financial compensation should not be entirely excluded. There should not be any compensation relating to training of an AI System with a vast amount of un-curated material from sources that are not protected with machine-readable reservations or technical measures.

   If a new special Exception is introduced for the Use of Copyrighted Data to Train AI Systems using for example a collective licensing model, there could be some sort of compensation to the Copyright holders, .

   While a collective licensing model would enable compensation for authors and rightsholders without separate license agreements for the utilization of each Copyright-protected work, it could be introduced only to very specific AI System

training with Copyrighted Data owned by a restricted number of authors. Excluding the works of these authors should be easy, and it would be desirable to link such compensations to transparency obligations such as those discussed in question 20 below.

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.

NO. Use of Copyrighted Data to Train AI Systems is a technical process that does not endanger the author's personal connection to the work, and it does not contradict good manners and practices. For this reason, Use of Copyrighted Data to Train AI Systems as such does not normally infringe Moral Rights. However, there should not be any express exclusions of Moral Rights relating to Use AI systems or use of Copyrighted Data to Train AI Systems. An AI system is a tool that may be trained to or instructed to perform various functions, and it is not desirable to create Exceptions that would enable anyone to intentionally infringe Moral Rights.

Moral Rights should be a ground to object output if right to integrity or paternity is infringed.

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?

YES. It should be disclosed whether the training has been done with curated data or un-curated data. A curated dataset should be described in a manner that reveals the sources and the time when the curated dataset has been accessed/downloaded or used for training AI Systems.

An un-curated dataset should be described in a manner that reveals its source and the time limits for obtaining the material. Potential restrictions relating to geography or language should also be described.

If the training of an AI system involves material that is created for the purpose of training, a general remark thereof is sufficient. The time of updating the corpus of training material should be disclosed whenever the changes are substantial. The list of commercially available licensed databases should be listed if used.

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 <u>without</u> the consent of the Copyright holder.

Infringing items

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

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

    YES. There is no need to deviate from the general principles of copyright law, under which the essential elements (those elements that meet the originality threshold) of the original work enjoy copyright protection.

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

    NO. There is no need to deviate from the general principles of copyright law, under which copyright protects the expression of ideas, not the ideas or styles themselves. Such a deviation could unnecessarily limit the freedom of expression.

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

    NO. The fact that an AI System was trained on infringing material should not automatically make all its outputs infringing. Such a limitation could hinder technological development if the entire AI System would need to be retrained.

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

    NO.

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

    YES, provided that it involves the reproduction of copyrighted works or making them available to the public, and the Exceptions described above do not apply (see question 17 above).

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.

   YES. The Provider of an AI System should be liable for Copyright infringement, if the AI System itself is or causes the infringement by reproducing Copyrighted Data and the Provider of an AI System has caused the infringement (causality requirement). For example, this situation could arise where the Provider of an AI System uses only limited amount of Copyrighted Data to Train the AI System leading to the AI System reproducing the Copyrighted Data as an output.

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

   YES. The user who exploits commercially the AI System should be held liable for Copyright infringement, if the user has no valid license to use the AI System, or, causes the Copyright infringement, e.g., by distributing output that qualifies as an unauthorized reproduction of Copyrighted Data.

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

   YES. The final user of output generated by the AI System should be held liable for Copyright infringement if the final user causes the Copyright infringement, e.g., by providing a prompt that causes the unauthorized reproduction of Copyrighted Data. In this context, infringement should be evaluated in accordance with the general principles of currently applicable Copyright law, and therefore, infringement should necessitate that the output reproduces essential elements of the original work and, e.g., stylistic characteristics resembling the original work should not alone constitute a Copyright infringement.

   The consequences of the infringement should depend on whether the final user has acted wilfully or negligently. Reasonable compensation should be available to the Copyright holder for the use of Copyrighted work irrespective of the final user's good faith or awareness of the infringement.

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

   YES. Any other person that causes a Copyright infringement, such as a distributor distributing an infringing output generated by the AI System, should be liable for the infringement.

Sanctions

24) What <u>sanctions</u> 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 <u>Copyrighted Data</u> used to train AI System still present in the dataset or in the AI System? Please answer YES or NO. Please explain if needed.

   YES. It should be possible for a competent court to order the destruction of Copyrighted Data used to train an AI System without authorization or applicable Exception, if the infringed work is still present in the dataset or in the AI System, and separable from the AI System itself. In addition, the court should be able to prohibit the infringer from proceeding with or repeating the infringing act (i.e., training of the AI System with infringing material).

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

   YES. A competent court should be able to order the destruction of an AI System or issue an injunction regarding the use of the AI System, if the AI System itself infringes Copyrighted Data, e.g., if the source code of the AI System is a copy of Copyrighted Data. However, the threshold for these measures should be high in relation to the entire AI System, and the measures should not be available solely because Copyrighted Data has been used to train the AI System, if the AI System itself does not infringe any Copyrighted Data (in which situation reasonable compensation and damages would still be available for the Copyright holder whose Copyrighted Data has been used to train the AI System).

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

   YES. It should be possible for a competent court to order the destruction or injunction regarding use of outputs infringing Copyrighted Data. In addition, recall from commercial channels should be available.

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

   YES. The award of damages should be available while punitive damages are not necessary and harmonisation to include punitive damages is not desirable.

The framework of damages should be left to be decided on the national level as there is a need to maintain coherence in the sanctions.

A competent court should be able to order a Copyright infringer to pay reasonable compensation to the Copyright holder. If Copyright is infringed wilfully or negligently or by committing a Copyright offence or violation, the court should also be able to order the infringer, in addition to the reasonable compensation, pay damages to the Copyright holder for all other losses.

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

NO. Harmonization of confiscation of assets to the state is not desirable. However, profits generated by the operation of the AI System should be available as damages (loss of profit) to the Copyright holder. In Copyright offence cases, it should also be possible to confiscate financial benefit gained by the offence.

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

YES. The most serious and repeated Copyright offences should be sanctioned under criminal law. There is no need to harmonize criminal law.

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

25) Please explain, in the following cases, if and under what conditions <u>outputs</u> 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 <u>Exception</u>. Please answer YES or NO. Please explain if needed.

YES. if the Exceptions related to the use of Copyrighted Data to Train an AI System does not extend to allowing outputs which reproduce the Copyrighted Data.

Reproducing the work and making copies of the work available to the public in whole or in part, directly or indirectly, temporarily or permanently and by any means or in any form should constitute a Copyright infringement unless there are Exceptions that allow such acts.

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

YES. The fact that Copyrighted Data has not been used to train the AI System should not automatically make all its outputs lawful. The output can still be considered infringing if the general criterion for finding infringement is met. An AI System is a tool that can be used to create reproductions.

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

NO. the infringement should be considered under basic principles and general criteria for finding infringement.

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.

YES. The Provider of an AI System should be held liable for Copyright infringement, if it has instructed the AI System to make unauthorized reproductions of works.

A Provider of an AI System is often in a position where it has no control or visibility of user's actions and cannot monitor the use of its AI System. A Provider of an AI System can apply technical restrictions that prevent output that contains direct reproductions from the source material. In addition, a Provider of an AI System is in a good position to either mitigate or assume contractual risks.

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

YES. The user who exploits commercially the AI System should be held liable for Copyright infringement, if the user causes the Copyright infringement, e.g. by distributing output that qualifies as an unauthorized reproduction of Copyrighted Data.

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

YES. The final user of the AI System should be held liable for Copyright infringement, e.g., if the final user provides a prompt that causes the unauthorized reproduction of Copyrighted Data. Furthermore, final users of output generated by the AI System should also be held liable if they, e.g. make available to the public outputs considered as infringing. The consequences of the infringement should depend on whether the final user has acted wilfully or negligently.

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

YES. Any other party that causes a Copyright infringement, such as a distributor distributing an infringing output generated by the AI System, should also be liable.

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.

There are many open questions relating to training and use of AI Systems. In areas where the use of AI Systems enables free-riding at the expense of the livelihood of people, it would be desirable to consider long term consequences when assessing whether Copyright or some other law regime allows prohibitions, restrictions or remedies. Functions and creativity that cannot be entirely replaced should be protected to avoid long term losses. AI Systems should contribute to efficiency and well-being and it should remain a tool for creativity.

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

N/A.