Response to WIPO Conversation on Intellectual Property and Frontier Technologies (Fourth Session)

Dr Luo Li

Assistant Professor in Law

Coventry Law School, Coventry University, United Kingdom

Introduction

Since artificial intelligence (AI) and its applications extend to innovative and creative areas such as generating inventions, literature outputs, painting and art design, there has been a series of discussions and concerns on AI-related ethics and legal issues as well as relevant impacts and implications to the relevant parties and the industries themselves. Considering the applications of AI and AI-related legal issues refer largely to innovative and creative areas as well as intellectual property (IP) concerns, the World Intellectual Property Organisation established the WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence since 2019 to bring stakeholders from diversified sectors to discuss the implications and impact of AI on IP policies and legislations. Until now, WIPO has held three sessions of the WIPO Conversation on IP and AI.[1] More importantly, WIPO builds the Issues Paper on IP policy and AI which lists a series of issues in detail for discussion purposes. This includes the areas of patent, copyright, data, design, trademarks and trade secrets, as well as technology gap, capacity building and accountability for IP administrative decisions.

[1] WIPO held the WIPO Conversation on IP and AI (first session) on 27 September 2019 to discuss and formulate an AI-related IP question list. Then WIPO published a draft Issues Paper on IP Policy and AI for the public consultation. Later, WIPO held the second session of WIPO Conversation with a discussion of the revised Issues Paper on 7–9 July 2020 and the third session on 4 November 2021.

It is admitted that the success of AI relies on three pillars which are respectively data, computational power and algorithms. Digitalisation and powerful cyber networks build perfect big data for AI training and generating purpose. An AI without full support for massive data is nothing at all. Therefore, data collection, data production, data refinement, use and its protection would influent AI investment. This would also raise a series of IP-related issues. In this case, WIPO organises and will hold the fourth session, WIPO Conversation on Intellectual Property and Frontier Technologies on 22-23 September 2021, with the focus on data beyond AI in a fully interconnected world and data in the current IP system.[2]

This report responds the issue 8 proposed by the WIPO Revised Issue Paper[3] provided at the second session of the WIPO Conversation on IP and AI in regard to the use of the data for the purpose of AI training and machine learning purpose.

Issue 8: Infringement and Exceptions

24. An AI application can generate creative works by learning from data with AI techniques such as machine learning. The data used for training the AI application may represent creative works that are subject to copyright (see also Issue 11). A number of issues arise in this regard, specifically,

(i) Should the use of the data subsisting in copyright works without authorization for machine learning constitute an infringement of copyright?

(ii) If the use of the data subsisting in copyright works without authorization for machine learning is considered to constitute an infringement of copyright, what would be the impact on the development of AI and on the free flow of data to improve innovation in AI?

(iii) If the use of data subsisting in copyright works without authorization for machine learning is considered to constitute and infringement of copyright, should an explicit exception be made under copyright law or other relevant laws for the use of such data to train AI applications?

(iv) If the use of the data subsisting in copyright works without authorization for machine learning is considered to constitute an infringement of copyright, should an exception be made

[2] WIPO Conversation on Intellectual Property and Frontier Technologies: Provisional Agenda (2021) WIPO/IP/CONV/GE/21/INF/1/PROV, 2–4.<https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_conv_ge_21/wipo_ip_conv_ge_21_inf_1_prov.docx> accessed 20 August 2021.

[3] Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence (2020) WIPO/IP/AI/2/GE/20/1 REV. <https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1_rev.docx> accessed 24 August 2021.

for at least certain acts for limited purposes, such as the use in non-commercial user-generated works or the use for research?

(v) If the use of the data subsisting of copyright works without authorization for machine learning is considered to constitute an infringement of copyright, how would existing exceptions for text and data mining interact with such infringement?

(vi) Would any policy intervention be necessary to facilitate licensing if the unauthorized use of data subsisting in copyright works for machine learning is considered an infringement of copyright? Would the establishment of mandatory collective management societies facilitate this? Should remedies for infringement be limited to equitable remuneration?

(vii) How would the unauthorized use of data subsisting in copyright works for machine learning be detected and enforced, in particular when a large number of copyright works are created by AI? Should regulations require logs of training data to be recorded?

(viii) If an AI application autonomously generates a work similar to an original work contained in the data used to train the AI application, would this constitute copying and hence infringement? If so, who would be the infringer?

Data related Terms

The current AI system is based on statistical learning and analysis, which means its learning and generating capability relies on massive data training and support. A sufficient quantity, high-quality and well-annotated data determines the success of all machine learning projects. Machine learning refers to 'a branch of AI and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy'.[4] Data mining relies on human intervention to extract information whereas machine learning could teach itself to improve the behaviours and adjust accuracy. Machine learning also utilises data mining techniques to improve the algorithms. From this point of view, the data mining technique plays the role of an input source for machine learning and a foundation role for AI.

In the AI training process, research and technique on text and data mining are quite necessary. In the official website of the Intellectual Property Office of the United Kingdom (UK), the term "Text and Data Mining" (TDM) is commonly used to describe 'the use of automated analytical techniques to analyse text and data for patterns, trends and other useful information'.[5] The

[4] IBM Cloud Education, 'Machine Learning' (*IBM Cloud*, 15 July 2020) <https://www.ibm.com/cloud/learn/machine-learning> accessed 2 September 2021.
[5] UKIPO, 'Exceptions to Copyright' (2014) Gov.UK <https://www.gov.uk/guidance/exceptions-to-copyright#text-and-data-mining-for-non-commercial-research> accessed 23 August 2021.

Copyright in the Digital Single Market Directive (DSM Directive) in Europe defines this term TDM as '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'.[6] In fact, text mining and data mining have common features and different features. Data mining means 'the computational process of discovering and extracting knowledge from structured data'.[7] Text mining is 'is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights'.[8] As the most common data type, text data is divided into three categories: structured data, unstructured data and semi-structured data.[9] Most data in the world stays in an unstructured format, which covers text from social media, product reviews, video and audio files for example.[10] Therefore, text mining research would be very valuable because it could transform these unstructured documents into a structured format for analysis purposes.[11] No matter data mining or text mining, 'it works by copying large quantities of material, extracting the data, and recombining it to identify patterns.'[12]

However, those large quantities of materials for text and data mining research as well as machine learnings and training are largely from copyrighted materials. In a copyright context, any use of a copyrighted work needs the prior consent of the copyright owner except for regulated limited situations and exceptions. Otherwise, those unauthorised use would be treated as copyright infringement. Back to the AI training situation, to answer the question of whether such copying and use of copyrighted works for text and data mining and machine learning purposes would constitute a possible copyright infringement or would be treated as exceptions, it is necessary to analyse the nature of how the data has been used in the text and data mining and machine learning process and what the protection scope of copyright law would be.

---

[6] Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, art 2(2).
[7] Cambridge Libraries, 'Text & Data Mining: What is TDM?' (*Cambridge LibGuides*, 28 June 2019) <https://libguides.cam.ac.uk/tdm/definitions> accessed 28 August 2021.
[8] IBM Cloud Education, 'Text Mining' (*IBM Cloud*, 16 November 2020) <https://www.ibm.com/cloud/learn/text-mining> accessed 23 August 2021.
[9] Ibid.
[10] Ibid.
[11] Ibid.
[12] UK National Archives, 'Text Mining and Data Analytics in Call Evidence Responses' (2014) <https://webarchive.nationalarchives.gov.uk/ukgwa/20140603093549/http://www.ipo.gov.uk/ipreview-doc-t.pdf> accessed 2 September 2021.

Use of the Data

It can be seen that data is normally used in several ways during the text and data mining and the machine learning and training process.

Information Extraction

As text and data mining technique are to explore those undermined relationships in those structured and unstructured data, one of the important functions for this "mining" technique is information extraction (or called data extraction) so as to find out relevant patterns, trends and correlations. For the purpose of achieving data extraction, researchers need to copy and input a large number of data which would involve massive copyrightable materials. Therefore, there are two issues that need to be analysed and clarified: firstly, how to identify the actions of data copying and inputting made by researchers from a copyright perspective; secondly, what is the nature of the action of "data extraction" in the copyright context.

A universal copyright principle recognised in the world is that copyright law only protects expressions of ideas rather than the idea itself. This means copyright law protects those original expressions that underline the works but not apply to facts, ideas, procedures, methods, etc. In this case, if any data relates to non-original expressions, it can be freely used for text and data mining as well as machine learning and training without copyright infringement issues.

If the researchers lawfully access and read the data, it would be no risk of copyright infringement. However, if they copy a whole or substantial part of copyrightable works for text and data mining/machine learning and training, it might be a risk of involving infringement. The good thing is most copyright laws at the national level provide limitations and exceptions allowing the public use. That is to say, even if the data involves copyrighted works, the data can be used as long as it falls into exceptional situations. In most national copyright laws, those non-commercial research and private study are normally covered by copyright exceptions. Although some countries such as the Copyright, Design and Patent Act (CDPA) in the UK copyright law and EU's DSM Directive expressly allow the text and data mining as an

exception,[13] this exception is strictly limited. For example, the CDPA only allows making a copy of a copyrighted work for text and data analysis with the purpose of non-commercial research whereas the DSM Directive requires text and data mining for the purpose of scientific research only. In practice, both commercial and non-commercial entities have been engaged in machine learning and AI training sectors. It is difficult to point out certain text and data mining research or machine training would be purely scientific and/or non-commercial, considering there is a possibility of shifting from a non-commercial starting point to a commercial achievement. Therefore, such a conservative approach to text and data mining as well as machine training data perhaps would be an issue for the existing copyright system.

Another issue is about the nature of "data extraction". Data extraction is the first step for text and data mining as well as machine learning and training. Data extraction refers to 'the process of collecting or retrieving disparate types of data from a variety of sources, many of which may be poorly organized or completely unstructured'.[14] While data extraction requires collecting, checking and temporary copying data for retrieving and extracting purposes, its application seems to be different from a normal collecting and copying purpose. Data extraction is not aiming at reproducing "expressions" of copyrightable materials but to extract informational value through these copyrightable materials. In other words, for example, approaching the data extraction method to literary works would only allow a text extract with annotation of labels such as sentiment tags, named entities and addresses, it would not refer to a replication of the literary works themselves. From this point of view, data extraction does not fall into copying "expressions" of the copyrightable materials but more like to dig and access ideas and facts rooted in these materials. Besides, the purpose of data extraction is not pretending to compete with authors of copyrighted literary works in the market. For instance, the text extraction method is used for machine natural language training so as to achieve automatic translation rather than reproducing copies of literary works via the extraction method.

Therefore, treating data extraction as a pre-potential-infringement nature is not appropriate and counting text and data mining as a copyright exception would be weird as well.

[13] Copyright, Design and Patent Act 1988, s 9A; DSM Directive, art 2(2), art 3–4 and 7.
[14] Talend, 'What is Data Extraction? Definitions and Examples' (*Talend*) <https://www.talend.com/resources/data-extraction-defined/> accessed 26 August 2021.

Information Storage

Text and data mining and machine learning require data extraction, transformation and loading process, which is normally called ETL. During this process, those high-quality, extracted and refined data are then delivered to a data warehouse for storage and analysis. Therefore, text and data mining and machine learning often involve both temporary and permanent copies of copyrightable materials. Normally, data would be permanently stored for the data set preparation. Data set refers to 'a file that contains one or more records'[15] which are 'the basic unit of information used by a program running on z/OS'.[16] In other words, a data set is a collection of data. Researchers may also temporarily store data for the analysis of the data set. Therefore, both temporary storage and permanent storage would lead to copying of data (it may cover both raw data and copyrighted materials such as literary, music and artistic contents). Copying raw data would not have any copyright issue but copying those copyrighted materials would be treated as infringing exclusive rights that copyright owners enjoy if such copying is unauthorised.

Nevertheless, it is worthy consider whether such action of "storage" for text and data mining in the purpose of machine learning and training should be determined as the action of "reproduction". The machine-learning-purposed data storage much likes to store memories (the data) into machine/AI's neural network. While machines use those stored data for training, self-analysis and adjustment, it simply uses their remembered (stored) information (or copyrighted materials). In this case, the memories (data) can be concrete information; can be an abstract form (creative elements of copyrighted materials). While humans remembering a novel or painting would not be treated as infringing copyright, AI/machines remembering (storing) this novel or painting is determined as an infringement. In this case, it is necessary to review the action of "storage" in a machine-learning-purpose and distinguishing from the traditional way of digital storage in computers, as well as consider a possible expansion of the interpretation to the word "storage" from a digital copyright perspective.

Producing Output via Data Input

[15] IBM, 'Zosbasics' (*IBM*) <https://www.ibm.com/docs/en/zos-basic-skills?topic=more-what-is-data-set> accessed 28 August 2021.
[16] Ibid.

While those data embracing copyrighted materials are used for the purpose of AI producing outputs, it looks like it is easy to identify a potential copyright infringement. This is because AI-produced outputs links closely with copyrighted materials – AI system remembers the data embracing such materials during the training and then AI system can produce new outputs based on their remembered data. Although these AI-produced outputs would not be the same as that of copyrighted materials, such "producing" would still be treated as making copies (a substantial part of copies) or be determined as communicating to the public. Making copies or communicating to the public are exclusive rights of copyright holders. Therefore, if any human make such a copy or communicate it to the public, this would be treated as copyright infringement. In this case, AI "producing" outputs should be treated as the same.

However, one of the issues about AI-produced outputs is that it seems to be not clear about the copyright authorship and the ownership. In this case, it would be difficult to identify who should be the appropriate party taking the liability for AI copyright infringement. After all, the producing process for the AI-produced outputs is autonomously rather than manually. Therefore, addressing the authorship and ownership issue in the AI-produced outputs is extremely necessary for the purpose of answering the liability issue in copyright infringement.

Conclusion

Technology does not only change the ways of humans lifestyle and bring challenges to the existing legal system, but it also offers human to consider a broader picture and revisit existing value to this world. While a new way of the use of intangible assets (here point at data, copyrighted materials, etc.) appears due to the technology change, it is necessary for us to identify its technical changes leading such new way of use as well as the nature of such use in both the technology context and the existing legal context. This is because it would make us understand more accurately the substantial features of such new use as well as achieve an appropriate legal response.

For the issue of using data for AI training purposes, we shall not limit our views only in an existing copyright context since such a copyright system is designed only for human use perspective. To answer the question of whether such use of data relates to copyright infringement or something else, the first thing that we need to understand is how the data is

applied or used step-by-step during the whole AI training and producing period in a technical context. Then we need to consider whether we shall apply the personification approach to determine machines using data is similar to human using data, or we shall create a new way of approach to determine the difference. In this case, the topic on the use of data and rights in data should perhaps beyond the copyright system and embrace other areas such as technology, competition, privacy and ethics.