AJPark

Copyright

Copyright risks associated with generative artificial intelligence

With the availability and popularity of large language model (LLM) artificial intelligence (AI) systems such as ChatGPT, legal issues are now more likely to arise than with the use of AI for pure scientific or statistical research, let alone autonomous robots, cars and lawn mowers.

For example, when using new generative AI there is a possibility that the output could include parts of copyright protected text or images which were included in the data inputted during the AI learning phase. Furthermore, such "data" could also include music in the form of sound recordings, for example. These all might constitute copyright works and unless their partial inclusion in generated AI outputs was authorised by the copyright owner such outputs may potentially constitute copyright infringement.

While other laws may also protect website data against certain usages this information sheet focusses on copyright law. Like human intelligence, AI must be provided with and memorise information on topics which it can then subsequently use when required to provide potential solutions in response to specific questions, requests and prompts put to it. There is a possibility for copyright issues to arise from both inputting 'training' materials to AI and its generation of 'answers'.

The new AI has caused copyright issues to arise for (i) creators of the materials which are used for teaching an AI system, (ii) AI companies and (iii) the latter's users.

Copyright works and copyright infringement

Copyright is a legal right (a property right) given to 'creators' such as authors and artists. For there to be a copyright issue in relation to AI some of the data collected and input into AI must constitute copyright 'works'. Those that are relevant to AI include literary works (texts and even computer programs), artistic works (images including photographs), musical works, films and sound recordings.

However, data which is pure information does not normally attract copyright and nor do ideas as opposed to the creative expression of an idea. Where 'data' is a copyright work and not just information, the copyright in it will be infringed by making an unauthorised copy of it. However, using data from open datasets will avoid potential copyright issues because its use is licensed (authorised) by the data providers.

Copyright is infringed by carrying out without permission one or more of the exclusive acts restricted to copyright owners. Two of those particularly relevant to AI training and AI outputs are (i) copying a work and (ii) communicating (transmitting) a work to the public over the internet

AI inputs and outputs

Collecting LLM AI learning data

Apart from tedious manual data extraction from web sites the following three techniques and sources may be commonly used with each presenting greater or lesser risk of copyright infringement.

Web scraping

Using readily available software, including scraping software provided as apps in AIs such as ChatGPT. Scraping web pages results in copying and storing text, images, sound recordings from web pages.

Open-source datasets

There are many open-source datasets available on the internet covering a multitude of topics and being 'open-source' are usually free of copyright problems. For example, Kaggle.

Synthetic datasets

These datasets are generated using computer programs rather than extraction from real world data. While they are useful when real world data is difficult to obtain, an additional benefit is that the data they contain will not have been acquired by simple copying.

AI generated outputs

The potential copyright issue with AI outputs is that they may contain, for example, text which includes passages literally taken from input texts stored in the AI memory. Under New Zealand (NZ) law it could be an infringement of copyright if such text passages while not a complete copy are a copy of a substantial part of any input text.

Copyright cases

While NZ courts have not yet considered AI copyright matters, cases have been launched in the United States (US). An example of a case where millions of photographs have been copied off the net into an AI system is the 2023 US case brought by Getty Images against Stability AI who has used them to teach the Stable Diffuse AI art generation system which has generated outputs incorporating some of the Getty photographs.

In 2023 there has also been an AI generated song launched on TikTok and Spotify which sounded like a Drake and The Weeknd song and attracted millions of listeners. Universal Music Group persuaded the streamers to remove the song, but a US court has yet to decide on whether the song infringed copyright or any other law. An important copyright issue will be whether a song which mimics artists voices, lyrics and musical styles amounts to an infringement of copyright.

A website owner will often not be the owner of any copyright in texts and images contained in its web pages which have been scraped to form AI inputs. Such copyright may be owned by companies or individuals who have limited ability to finance legal actions.

Reducing the risk of Court cases

Scraped materials may constitute pages from online journals and magazines. New Scientist is one of them and has argued that copyright is failing them because like their regular readers private AI scrapers should at least have to pay a subscription to increase the odds that the publisher may withhold commencing court proceedings to enforce copyright.

AI and NZ Copyright

The purpose of having AI generate a work, such as a text, an image or music may also be relevant in determining if there has been copyright infringement. However, NZ copyright law provides less exceptions to acts which could constitute copyright infringement than the US, for example. The US has a significant exception to copyright infringement which is the doctrine of 'fair use'. Under this doctrine a copier will have a defence if, for example, the copies they make are not made for commercial purposes and/or if they are 'transformational' versions of the original work.

Therefore, in the US, web page scraping for AI teaching data could be considered fair use and AI generation outputs which simply incorporate the 'style' of one or more inputs might be considered as transformations of them and not copies. To the contrary, NZ simply has somewhat restricted 'fair dealing' exceptions. For example, where the copying of a work is purely for the purposes of research or private study.

However, the NZ Copyright Act does have similar remedies to those provided under the US Digital Millennium Copyright Act in that where a streamer who falls within the definition of an 'internet service provider' stores, say, a pirated sound recording then they must delete it as soon as they become aware it is infringing to avoid themselves becoming a copyright infringer. This remedy is available even before a case against the streamer is brought to a court.