Kluwer Copyright Blog

Reconceptualizing the Reproduction Right in the age of AI

Rita Matulionyte (Macquarie Law School) · Monday, March 10th, 2025

Image via Pixabay

We have so far seen a considerable (and increasing) discussion on AI and copyright infringement, especially in terms of how current exceptions such as TDM and fair use apply and whether new exceptions or remuneration models are needed. One question on which there has been little discussion is whether the reproduction right is triggered when an AI module is trained using content protected by copyright. It is essential to answer this question before we discuss defences to copyright infringement claims or remuneration for the use of protected content in the AI learning process. If the AI learning process does not result in (non-incidental) copies of works used during AI training, then we do not need to discuss copyright exceptions, and we cannot talk about remuneration of right holders at all.

What/where is the legal problem?

It is clear that when protected content is used in the AI training process, the reproduction right is triggered at various stages. First, it could be triggered when compiling content (texts, songs, etc) into training datasets, though not in all cases (e.g, some training datasets contain only hyperlinks to

content stored online rather than copies of content). However, while this could mean that data creators/compilers could be liable for direct unauthorised reproduction, AI developers (i.e. those who train the module) often do not compile datasets themselves but use pre-existing training datasets. In such a case, AI developers might argue that they are not making copies of training data themselves, at least not permanent or lasting copies (though they might make at least temporary copies of pre-existing datasets as part of the training process, which could be covered by fair use or temporary copying exceptions available in many jurisdictions).

Second, the reproduction right could be triggered at an output stage if the AI output reproduces a substantial part from the work contained in the training dataset. However, in current Generative AI models this happens relatively seldom (an instance of 'memorisation') and AI companies are working to minimize or eliminate such 'memorisation' instances entirely to avoid copyright infringement allegations with relation to AI outputs.

The most controversial question is whether reproduction occurs at the stage of AI training, i.e. when the AI module is iteratively exposed to training data which change and eventually determine the parameters of the algorithm. AI developers and some scholars, especially in the US, argue that this training process does not entail or lead to copying. Arguably, AI models do not store copies of training data. They merely 'ingest' them or 'learn' content into their parameters. These commentators argue that models merely process text or other content using tokens (common sequences of characters) and learn the statistical relationships between tokens. They conclude that the use of copyrighted data in AI training does not trigger the reproduction right. Another group of scholars suggest that perhaps the AI training process entails some sort of temporary copying, but these copies are most likely covered by exceptions, e.g. fair use or temporary copying.[1]

Proposal: expansive interpretation of the reproduction right

What I argue is that even if AI modules indeed do not store copies of training data, the so-called 'ingestion' or 'learning' of the protected content into AI modules should be considered as equivalent to copying of this content. Perhaps it is true that AI modules do not store copies of all works on which they were trained in a manner with which we are familiar (e.g. in files and folders). Rather, the content is disintegrated and 'ingested' in the parameters of the algorithm in a new manner which we have not seen before. However, I argue that this 'ingestion' or 'learning' of content and its 'embedding' in algorithmic parameters is a new form of storing of the content. Even if in a disintegrated manner, the content has been embedded in the model, in its parameters, for certain reuse purposes. Namely, it will be used to create new content. Sometimes this new content (AI outputs) is entirely different from the individual pieces of content on which it was trained, in other cases it might be very similar to the content in its training dataset – both in terms of ideas and expression. It will depend on how AI developers designed and trained the algorithm. While GPT4, which is a module behind ChatGPT, is designed in a way to avoid outputs similar to the training dataset, the Next Rembrandt project was designed to output artworks that are very similar in style to Rembrandt paintings on which it was trained.

Rationale of this proposal

There are at least three reasons that support the extension of the reproduction right to the use of

protected content in the AI learning process (and embedding that content in AI models). First, historically, the rights of authors, publishers and other right holders have been expanding with the emergence of new technology: from reprinting they have expanded to translation, performance, broadcasting, and communication over the internet; from covering only analogue copies the reproduction right has expanded to digital copies (essentially 1s and 0s). It was acknowledged that right holders have legitimate interests to control the use of their works in the context of new technologies. The use of works in AI training is another new type of use, resulting from new AI technology, and there are no good reasons why right holders should not be able to control these commercially valuable uses that have the potential to affect significantly their interests in various ways (including the possible loss of creative jobs).

Second, this interpretation of the reproduction right would enable authors to exercise their rights in the age of AI. Namely, only if it is recognized that reproduction is occurring during the AI training process, will they be able to license such use of their content and enforce their rights. If we conclude that AI training does not lead to copying, then we cannot talk about right holder remuneration (and do not need to talk about exceptions, opt-outs etc).

In addition, recognizing that AI models contain what is equivalent to copies of works would enable right holders to exercise their rights not only with respect to models trained in their own jurisdiction, but also with respect to AI models trained in another jurisdiction, as long as this AI model is offered in their jurisdiction. For example, the Stable Diffusion algorithm was trained in the US on billions of images protected by copyright. If we agree that the trained stable diffusion model contains copy-equivalents of these images, then offering this trained AI model in another jurisdiction (e.g. the UK or EU) could arguably amount to the public communication of these embedded copies to the public in the UK and would trigger the public communication right under the UK or EU copyright laws.

Finally, this solution would align with most recent international policy approaches. The EU AI Act (Article 53(1)(c)) already requires that all AI modules offered in the EU should respect EU copyright laws. The UK is currently considering a similar solution. Recognizing that AI modules contain what is equivalent to copies of training data would implement this general policy approach into the copyright framework and make it enforceable.

Final notes

It is important to emphasise that the revision of the reproduction (and public communication) rights to include the use of works in the AI training process is one, but not the only, measure needed to reset the balance in copyright law in the new age of AI. If we decide to expand the rights of right holders in this manner, we will also need to think how to (re)establish the balance of different interests involved. Some jurisdictions that do not yet have any exceptions that could apply in the context of AI will need to discuss if and what further exceptions are needed (e.g. Australia). Jurisdictions that already have certain exceptions (fair use or TDM) need to reevaluate their suitability in the AI context and many are currently working on this (e.g. the UK). Further, there will be (or is) a need for effective licensing and enforcement mechanisms to ensure that right holders can benefit from these rights and receive a part of the profit that AI industries generate. Finally, we also need measures to ensure that the revenue flowing from AI industries into creative industries is distributed equitably, i.e. that it does not stay in the hands of large secondary right

holders (e.g. large publishers, record companies), but reaches individual creators.

For a more in-depth discussion on this issue please check: Matulionyte, Rita, Reconceptualising the Reproduction Right in the Age of AI (December 02, 2024). Available at SSRN: https://ssrn.com/abstract=5041741 or http://dx.doi.org/10.2139/ssrn.5041741

---

[1] Eg Pamela Samuelson, 'Generative AI Meets Copyright' (2023) 381 *Science* 158, 159; Matthew Lindberg, 'Applying Current Copyright Law to Artificial Intelligence Image Generators in the Context of Anderson v. Stability AI, Ltd' (2024) 15 *Cybaris* 37, 60-61

To make sure you do not miss out on regular updates from the Kluwer Copyright Blog, please subscribe here.

This entry was posted on Monday, March 10th, 2025 at 9:37 am and is filed under Artificial Intelligence (AI), Infringement, Reproduction (right of)
You can follow any responses to this entry through the Comments (RSS) feed. You can leave a response, or trackback from your own site.