EIP

Sustainable models: content licensing for generative AI training

It is fair to say that no technology in recent times has created so much excitement and controversy as generative AI, and in particular Large Language Models (LLMs). Generative AI companies and their backers have their sights on generating billions, if not trillions of dollars, directly or indirectly from their models. Models which, in the vast majority of cases, are trained on original works without permission from, or payment to, their authors. The works used for training are primarily obtained by scraping the open web, a uniquely vast repository of publicly accessible knowledge and information, constantly updated to reflect the latest developments in human knowledge and culture.

Even leaving aside the big questions of what is ethical and what is equitable, it could be argued that the generative AI industry must eventually move from being extractive to being sustainable. The AI companies need a continuing supply original works so that their models can continue to reflect an ever-changing world. And as we move to a world where more and more information is intermediated via LLMs (i.e. "answer engines", rather than "search engines"), creators need a means of remuneration to enable them to continue creating and sharing their work, which is ultimately in everyone's interest.

These issues haven't escaped the scrutiny of the law. Thanks to a largely harmonised copyright system, creative works are by default protected by copyright wherever they are created and wherever they are subsequently used. But that harmony, and the copyright system itself, are now under threat, as different countries and regions, pursuing different policy goals, have started to diverge in their approach to generative AI. In the US – home ground of the most illustrious AI companies – battles are raging in the courts over what constitutes "Fair Use" of copyright works. In the EU, generative AI training has seemingly been grandfathered into a Text and Data Mining (TDM) exception that predates the

technology, enabling training on copyright works provided the rightsholder has not explicitly opted out. Publishers and creatives argue that the opt-out is unworkable, while
p2 a recent case in Germany (GEMA v OpenAI) suggests that generative AI training can fall outside the TDM exception altogether. Under UK law there is less leeway for use of copyright works, though in a landmark ruling (Getty Images v Stability AI), it was found that importing a model into the UK trained on copyright works was not an infringement. The government is now considering its approach, aiming to balance its ambition to become an AI leader with the interests of its significant creative and media sectors.

Due to the scale of datasets needed for generative AI training (at least in its current guise), it may be unrealistic for AI companies to negotiate licences with each individual rightsholder. This cries out for a collective licensing scheme and a functioning data marketplace, perhaps analogous to that of the music industry. Another analogy is music and video streaming, where content creators are paid per stream. Something similar could potentially work for AI models, though it would require technology for attributing value either to models or outputs at scale. In either case, the volumes of data typically used in AI training could make such solutions costly and challenging to administer.

Collective licensing solutions do not fit well with the current dynamic and automated nature of web-scraping. An alternative, or additional, solution might involve integrating payment into web-scraping itself. During this process, a web-scraper requests a webpage via a URL, loads the HTML of the page (including JavaScript and all elements), and extracts and saves the data it finds. A list of URLs from which data has been extracted will be generated, and this information could be used to trigger payments to an account associated with the URL. This method would however rely on the companies sending out the scrapers to report honestly and transparently on what they have done, something which would be difficult to enforce or police.

Alternatively, web pages could require payment for web-scraper access. For example, software implemented on a web server could detect that a request for content is from a web-scraper, and in response, a payment element may first be returned which, if agreed to, will allow the scraper access to the URL. A solution along these lines has recently been implemented by US infrastructure company Cloudflare, which sits behind around a quarter of all websites. This solution would require less transparency and honesty from the AI companies, but would require operators to implement additional technology on their servers. To the extent that AI training (rather than web-scraping per se) is the relevant act, such a solution would also need to determine that the purpose of the web-scraping is to train an AI model. This may be critical as website owners would still wish to enable web scrapers for indexing by search engines to access their servers unimpeded. In the UK, a proposed amendment to the Data Access and Usage Bill, which would have

legally required web-crawlers to be transparent about their purpose, was recently rejected. Innovative technical solutions around detecting and/or authenticating access by different web-crawlers may also be developed.

In our view, the most impactful approach to safeguarding creators' IP rights is likely to involve a combination of technical and legal solutions. To the extent that courts and governments across the world decide that companies should require permission to use (and hence pay for) copyrighted content to train AI models, there will need to be scalable and workable technical infrastructure to support this. There is innovation in this space. But any technical solutions will work best if backed by strong copyright law. This will ensure that, even where attempts to circumvent the technology are made, there will remain the option of asserting authors' IP rights in the courts.