LIDC

INTERNATIONAL LEAGUE OF COMPETITION LAW

Competition Law Association

British Group of the
Ligue Internationale du Droit de la Concurrence
(International League for Competition Law)

LIDC Conference 8 November 2024

Session 2A - The AI Services Hungry for Data

Who gets to access and use data, and the copyright vs antitrust conundrum it gives rise to.

Panel: Pinar Akman (UK Competition Appeal Tribunal/University of Leeds), Matthew Cope (UK IPO), Lindsay Lane KC (8 New Square Chambers), Cedric Manara (Google)
Chaired by Giles Parsons (Browne Jacobson)

Technical Background

Per Linsday Lane KC, for the purposes of IP, it is helpful to divide the AI model into two, input and output.

Input:

- In order to create an AI model, you need to develop some kind of neural architecture, and you need to train that architecture on a dataset
- This dataset needs to be massive, given that the bigger the dataset that you train the model on, the better the outcome
- You then train the model on the dataset, going through iterative phases in order to develop the model
- This data needs to come from somewhere, and the general starting point for most AI models is to use data which has been scraped from the internet which, in Lindsay's view, raises interesting questions on the permissibility of this practice, questions which vary between jurisdictions
- From the perspective of IP, training the AI model involves creating identical copies of the training dataset; which IP rightsholders would say is a straightforward case of copyright infringement (subject to jurisdiction and any applicable exceptions or defences)

Output:

- Output refers to what happens when the user enters some kind of prompt into the model and gets a synthetic output
- In addition to copyright, Lindsay raised the point that outputs may present issue of trademark infringement and passing off, as well as copyright infringement.

What do we know about these markets?

There has recently been a European Commission consultation, from which a policy brief has been published. The panel discussed that the consultation identified the extremely dynamic nature of the generative AI market.

The process of training the data should be split into two stages:

1. Pre-Training
    a. Uses lots of publicly accessible data (which maybe, in itself, the subject of copyright)
2. Fine-Tuning
    a. Operators tend to rely on proprietary data licenced from third parties due to a lack of high quality publicly accessible data

Pinar Akman noted that there arises a competition issue in that the large players have access to data from their own eco-systems, and they benefit from their already enormous processing and memory capabilities. She noted that given the enormous resources that training an AI model requires, this creates an externality that presents a barrier to entry, and thus poses a threat to competition.

Further, Matthew Cope added that larger companies are able to make use of partnerships with publishers (and data holders more broadly) in order to licence data for training. These licence agreements often contain exclusivity clauses, which creates a barrier for small players and reinforces the market power of incumbents.

Cedric Manara disagreed, arguing that Google trains on the open web like any other company and that everyone in the market has access to the same data (provided it is not behind a paywall). He also raised the point that, from a copyright point of view, it would be administratively unworkable to seek licences for every single piece of data, which, in Google's view, gives them a good argument for fair use.

To counter Matthew Cope's point, Cedric argued that to make the data free to all would be anticompetitive, and that it is not the agreements themselves, but developments within the sector that can pose a competition issue.

On the IP issue, Cedric admitted that AI models do copy the data, tools like Google Translate are prime examples of why copying is beneficial. This brought the panel onto the final key talking point.

What are we looking to incentivise?

The panel discussed the competing interests at stake in the AI market, discussing protecting intellectual property rights, promoting competition and encouraging the development of the artificial intelligence sector. The question was raised as to which should be incentivised? The panel discussed the issues surrounding which might benefit society at large the best?

Lindsay Lane raised the arguments for protecting and promoting intellectual property, including noting that AI does not train very well on its own synthetic data. As such, there is a need for a constant stream of new data to keep models up-to-date and useful. This requires the incentivisation of the creative sector to continue contributing to the data pool from which the models can be trained.