Fair Use in Data Mining and Machine Learning: A Comparative Study between Mainland China and Taiwan

Gi-Kuen Jacob Li

Research and development of artificial intelligence is in full swing (again). Since Data mining and machine learning are two essential methods to develop AI, both are employed in a variety of fields, which may both involve reproducing and adapting works that are protected by copyright law. AI that is designed for natural language processing (NLP) often rely on literary works as training materials to construct the corpus, ontologies, and semantics of languages, to enable machines to "understand" and process human language. For example, AI that is capable of generating literary works usually requires copyright protected works such as novels, articles, or poetry as training materials. Similarly, AI that is built to generate works such as music, pictures, graphs, audiovisual works, and sound recordings could all face the same issue.

Under the fair use doctrine of US copyright law, it seems transformative fair use is not generally applicable to the varieties of data mining and machine learning. In order to establish transformative fair use, the goal of copyright – to promote science and the arts – should be furthered by the use; in other words, it is important to examine whether the use, as stated in *Campbell v. Acuff-Rose Music, Inc.*, "adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message[.]" While some AI designs certainly have great potential to further the progress of science and the arts, others could be purely commercial and can be substitutes of the training materials.

Similar to the US, mainland China and Taiwan seek to invigorate the AI industry, therefore relevant laws and policies are critical to facilitate research and development. While US copyright law adopt a flexible fair use standard, mainland China and Taiwan have different legislation. Mainland China follows the Berne three-step test with enumerated limitations, whereas Taiwan blended US fair use factors and categorical illustrations of limitations into the Copyright Act. However, the enumerated limitations in mainland China might not apply to either data mining or machine learning. Likewise, it is also questionable that current fair use doctrine in Taiwan is applicable to all scenarios of data mining and machine learning. Thus, this research aims to explore the application of current fair use doctrine in the data mining and machine learning setting, and how jurisdictions set out to fine-tune the fair use doctrine.