Your World First

C/M/S/ Law . Tax

Artificial Intelligence

AI and data mining – questions of copyright

AI solutions need large datasets and, in some cases, substantial vocabulary and linguistic analysis when reading and interpreting huge volumes of text. In the AI teaching phase, developers use many copyright protected works from literature and science, as well as user-generated content on blogs and social media. This can lead to conflict with copyright owners. Data mining – the extraction of contents from a database – may also affect databases and infringe the rights of the database producers. This article deals with copyright exceptions in force in Europe and their possible implications for data mining and AI development, and the competitive position of European business.

The importance of data mining

Text or data mining is an automated analytics method, which analyses digital text and data to discover information, patterns, tendencies and correlations. It is fundamental in generating the robust and varied data sets that underpin machine learning and deep learning, the processes that constitute the bedrock of AI.

One of the most important types of text mining is natural language processing (NLP), which employs linguistic analysis to enable a machine to “read” and “interpret” text. Many industries already use NLP as it is essential for developing state-of-the-art customer service with Q&A systems, intelligent digital or virtual assistants, personal assistants, chatbots, voice assistants and e-mail auto reply systems. Self-service phone banking applications use the same technology.

NLP is also valuable in digital and online marketing, where they are deployed to understand customer sentiment and awareness by analysing social media posts and customer behaviour on online platforms. The resulting analysis is used to target advertisements and to personalise marketing offers. In the recruitment sector, AI and text mining enable the automated extraction of information from CVs and automated interview calls. Analysing the written and spoken words of job applicants can accelerate the recruitment process.

Competitive advantage

Research organisations and businesses often confront legal uncertainty about the extent to which they can mine content. Where no exception or limitation applies, undertaking such activity would require authorisation from the content owners and often involve payment of a licence fee. This restriction can make R&D and scientific work difficult, hindering innovation and putting European businesses and researchers at a competitive disadvantage compared with their counterparts in the US and Japan, where exceptions permitting test and data mining are considered by many commentators to be more flexible.

The much-debated new Digital Single Market copyright directive partially solves these copyright issues by setting out two exceptions in relation to data mining: a general exception subject to restrictions and a specific exception for the non-profit sector. In certain circumstances, any AI developer is entitled to use copyrighted works for data mining purposes with some limitations and without any license or royalty payment.

New copyright exceptions

The new EU copyright directive sets out to resolve this copyright issue by creating two exceptions, a strong exception for the "non-profit" sector and a weaker exception for profit-making purposes.

A strong mandatory exception applies to reproductions and extractions by research organisations in order to carry out text and data mining of content to which they have lawful access for the purposes of scientific research. This exception is available to universities, research institutes and other organisations conducting scientific research, but only where the research is not-for-profit or where any profits are reinvested in research. This exception overrides any contractual provisions to the contrary.

By way of contrast, the second, weaker exception applies to any lawfully available material, without restriction as to profit motive. This exception only applies "on condition that the use … has not been expressly reserved by … right holders in an appropriate manner, such as machine-readable means in the case of content made publicly available online." In practice, this means that the existing regime, requiring a rightsholder's approval and, possibly, a licence fee, is likely to endure.

A real solution?

Although the European Commission recognises the importance of data mining and AI when it proposed the new copyright exception, many commentators argue that the final form of the exceptions is not an adequate response to the competitive challenges that EU businesses face. In particular, the strong exception excludes for-profit organisations, research centres, private for-profit universities, innovative firms and journalists.

In the US, by contrast, text and data mining are covered by the 'fair use' principle, regardless of whether the organisation concerned is for-profit or non-profit. The US courts have explicitly upheld text and data mining as fair use in several cases. Most famously, Google was able to invoke fair use when scanning books in the collections of its partner libraries and incorporating the works in a searchable database that could be used by scholars and researchers.

Contacts

Katalin Horvath
Senior Associate
T +36 1 483 4897
E katalin.horvath@cms-cmno.com

Further information, including a list of our offices, can be found at **cms.law**