Can the UK Become Competitive on Text-and-Data Mining for AI?

Dr Anton Howes

January 2024

- There is currently a great deal of legal uncertainty about commercial text-and-data mining in the UK
- Other countries have moved faster than the UK at securing the legal foundations for training AI
- Japan, Singapore and the EU each provide competitive models for the UK to consider

Data is the raw resource upon which AI is trained, through a process known as text-and-data mining (TDM). In order to ensure that AI is of a high quality, the data that underpins it needs to be plentiful.

In the UK there is currently a limited TDM exception from copyright, covering only copies made for the purposes of conducting non-commercial research that was introduced in 2014. So long as non-commercial AI researchers have lawful access to data – for example, if it is publicly viewable on the internet without needing to sign up to a contract to get past a paywall – they do not have to ask for a licence from the owners of the copyrighted material that is being mined.

Otherwise, in order to train AI, all other users have largely been expected to acquire licences from the people (or 'rights-holders') who own the copyright in the material that they wish to mine – in other words, the person who took a particular photograph, the writer of some text, and so on. In some cases, such licences may be given freely, but in general rights-holders expect to be remunerated for using their work for TDM. These licences can be extremely varied, as there can be any kind of agreement between two parties, with different price points for different kinds of licences too. Acquiring licences ultimately requires negotiation, though in some cases rights-holders will have signed up to collective licensing arrangements, by which a great many of them allow a particular entity to manage the negotiation of licences on their behalves.

That, at least, has been the expectation of many rights-holders. But some of the more established AI companies and other user groups interpret the law differently, and dispute whether the training of algorithms using lawfully accessed works constitutes copyright infringement.

Ultimately, this disagreement must either be decided in the courts or else pre-empted by a change in legislation in order to clarify the situation or to create new exceptions covering the use of TDM. And this looked, for a while, like it might have been impending.

In 2022 the Government announced, following a consultation, that it would implement a blanket exception from copyright for copies made for the purposes of TDM, without any opt-outs, and regardless of whether it is for commercial or non-commercial uses. In early 2023, however, after complaints from rights-holder groups, the Government said it no longer plans to extend the TDM copyright exception. Later that year, the Vallance Review recommended going ahead with extending the copyright exception for TDM to all purposes. The Government said it would be accepting the review's recommendations, but chose to interpret them as meaning that it should task the Intellectual Property Office with overseeing negotiations between users and rights-holders with the purpose of reaching a voluntary code of conduct that both sides could sign up to.

In the meantime, however, the continued lack of clarity has created a great deal of uncertainty – something damaging to both rights-holders and TDM users, and to the adoption of AI in the UK more broadly.

There is a serious risk that continued uncertainty will create a chilling effect on the adoption of AI services by consumers and other businesses at large, with implications for the productivity of the economy as whole: ordinary people must be able to confidently use AI services and the outputs that they generate without the risk that they will be held liable for unwitting copyright infringement because of the nature by which the AI was trained.

Thus far, copyright infringement suits have been initiated against AI companies in the USA. But the American system is very different to that of the UK. AI developers there have claimed that TDM on lawfully accessed works constitutes fair use – a loosely defined and principles-based legal concept that does not apply to the UK, where exceptions are instead strictly delineated.

Regardless of the outcomes of those US cases, however, the continued uncertainty in the UK is likely to encourage AI companies to relocate to jurisdictions that can offer greater certainty or which have more permissive exceptions. In Israel, for example, where there is a fair use system like that in the USA, the Ministry of Justice has attempted to provide certainty by issuing an opinion that copyright law does not prevent the use of lawfully accessed copyrighted material for text-and-data mining.

Otherwise, the most notable systems to compare to are those of Japan, Singapore, and the EU, each of which have approached this issue in very different ways.

Japan. The Japanese approach, introduced in its Copyright Act of 2018, is to have a broad exception for text-and-data mining. It permits copies of lawfully accessed material to be made without the need for licencing so long as the copies are not made for a human to enjoy the ideas or sentiments expressed in the work – that is, so long as only a machine is doing the reading, and the copy is made only for the purpose of training the AI, then the copy is permitted. This is not a blanket exception, however, as the law comes with two major safeguards for rights-holders.

One is that a copy made under the provision without a licence must not "unreasonably prejudice the interests" of the rights-holder whose work it is trained on – something to be determined on a case-by-case basis according to whether it conflicts with the current or potential market for the rights-holder's works. What exactly this means is yet to be seen.

The second is that a copy made under the provision without a licence cannot be used to create work from which we can feel the essential characteristics expressed in the work upon which it is trained. That is, it does not allow for the exception to be used so that an AI can, say, produce an image in a specific artist's style.

The Japanese approach thus creates many situations in which the need for licensing may be avoided entirely. But it still leaves significant areas in which licensing will still be needed: the creation of outputs that compete with the work upon which the AI is trained; and the imitation or sufficiently similar imitation of work upon which AI is trained. What is unclear is how far this extends.

Singapore. Singapore's approach, adopted in its Copyright Act of 2021, is to have a blanket exception for TDM. Singapore's law permits copies of lawfully accessed material to be made without the need for licencing so long as the copies are made for the sole purpose of computational data analysis – which it explicitly illustrates as including "the use of images to train a computer program to recognise images."

The exception also covers the preparation of material for the analysis, sending it onto others for the verification of results of the analysis, and collaborative research or study of the analysis. It even explicitly prevents the terms of use for the lawful access material from being used to override the exception through contract, rendering any such terms of use void.

Otherwise, the usual protections of copyright apply to AI-created outputs. Although the law permits AI to be trained on material without the need for licensing – it lifts all constraints on inputs to the training of AI – it does not permit the publication of work that is too substantially similar to protected work.

The European Union. The EU's approach, enshrined in the 2019 Copyright in the Digital Single Market Directive, is that the TDM of lawfully accessed material for any purpose – whether commercial or non-commercial – is permitted without the need for licensing. But rights-holders can explicitly opt out or "contract out" of allowing their work to be used for this purpose, so long as those opt-outs are done in a machine-readable way.

What is still unclear, however, is how well this system of opting out will work as it comes to be implemented by member states. The manner of opting out has not been standardised, and it is unclear how or whether someone would be able to prove the absence of a relevant opt-out when they trained AI upon the material. Nonetheless, the provisions allow for AI to be trained upon swathes of lawfully accessed material that rights-holders do not wish to commercially exploit through licensing arrangement – while preserving the ability of rights-holders to sell licences to AI companies that wish to licence their data. It thus allows for both the existence of a market in training data while also allowing AI to be trained upon unexploited material.

Dr Anton Howes is the Head of Innovation Research at The Entrepreneurs Network, a think tank for Britain's most ambitious entrepreneurs. Find out more about our work at tenentreprenuers.org.