Computer & Communications Industry Association
Open Markets. Open Systems. Open Networks.

AI and Copyright

With the Government launching a consultation on AI and copyright, this CCIA briefing is intended to answer questions policymakers might have and respond to some common misconceptions. It is also intended to provide a high level overview of some of the factors that the Government might need to consider in realising its objective of ensuring that “leading AI models” are trained in the UK.

What is text and data mining?

Text and data mining (TDM) is the process of analysing machine-read material. This is analogous to reading or listening to music as a first step in developing your literacy or cultural knowledge. It is often one step in gathering the data needed to train a leading AI model, but also includes other processes, e.g. developing classifiers to label often large amounts of unstructured data.

What does training an AI model mean?

To develop AI systems responsibly, making good predictions and supporting informed decision-making, an AI model must be "trained" on enormous quantities of data. This is one important form of TDM.

Balanced text and data mining exceptions within copyright law allow researchers, innovators and creators to use copyright-protected material under certain circumstances without permission from the copyright holder. The EU Directive on Copyright in the Digital Single Market has an exception for TDM in both commercial and noncommercial contexts. UK copyright law has an exception for "computational analysis" in noncommercial contexts (it does not use the term text and data mining.) One of the major issues in the consultation is whether to expand the UK exception to commercial contexts, something that has taken place in other jurisdictions such as the EU, Japan and Singapore. In the U.S., AI firms rely on the fair use doctrine to enable AI training.

Are outputs from AI models copies?

AI models and the training process can vary enormously, but the outputs are not copies. Generative AI systems use mathematical techniques to learn patterns and concepts as numerical parameters or weights and generate outputs based on that learning. This is a statistical process that is not copying:

- If I ask you to draw a picture of a cat, you will likely do so based on your knowledge of cats (e.g. they have whiskers) and your knowledge of how to draw — creating a new picture of a cat when prompted. This is closer to what generative AI models do.
- If I ask you to copy a picture of a cat, you will find a picture, look at it and try to match it as closely as possible. This is not what generative AI models do.

AI models take that learning process, but have access to a much greater volume of content available across the Internet. There is no particular piece of content that inspires a particular output, but a generative process reflecting a very large volume of content used to train the model.

This difference is important because it means that for foundation models, a lot of content which is read does not generate any particular value for which compensation is appropriate or practical. In the same way, people are not required to identify the full range of inspirations for their own work and compensate all those who have inspired them over the years. From a business perspective, any remuneration should (and generally does) focus on where specific content is particularly valuable for a particular set of use cases.

Will AI put creative professionals out of work?

AI applications complement human activity: coders who can work faster with AI suggestions and improvements; creators on video sharing platforms who could never afford to commission a custom theme song for their channel but can ask AI to help; video and photo editors with access to better tools to process their raw footage ready for editing. In other cases they meet needs that were rarely met by creative professionals in the past, stitching together short clips into home movies for example.

As has been seen with earlier waves of creative technology (e.g. digital editing software), any tool that improves productivity has the potential to reduce the need for some kinds of labour. But it will also increase the demand for other, often adjacent, kinds of labour. There are a number of reasons to think that the impacts in the short- to medium-term will not be falling demand for human creativity overall.

- The creative sectors as a whole are growing. While there is variation over time, DCMS statistics have generally found that the creative industries grow twice as fast as the regular economy. The most recent *Sky Is Rising* report by the Copia Institute and the CCIA Research Center looks at a broad range of trends and finds recorded music revenues, podcast listenership, number of scripted TV series, entertainment revenue, book revenues and video game revenues are all rising over time.

- The outputs of creative industries are often sold globally. If the UK's successful creative industries can work more efficiently, they can sell more in global markets. If not, they can lose market share.
- While people can be trained to take up many creative industry roles, there are probably limits on the availability of talent in some of the most successful creative industries. Easing those constraints by enabling people to work more efficiently allows the sector to grow and create other opportunities for complementary labour.
- AI tools reduce barriers to entry in creative fields, reducing the need for people (and smaller companies) to master some technical skills in order to bring their creativity to the market. This can enhance competition and innovation.

Can AI developers just avoid using copyrighted works?

As the requirements for copyright protection are so low, and the terms of protection are so long, the vast majority of content on the internet could be covered by copyright. Copyrights are not subject to registration (there is no database of in-copyright works which AI developers can check and rights holders have traditionally resisted this sort of transparency) so it is impossible for AI developers to know with certainty which content is protected by copyright (versus content for which copyright might have expired, for example, or which was exempted in the first place - e.g. facts). It should not be the task of AI developers to make decisions on the copyrighted status of works, due to the complexity of attributing copyright and the sheer amounts of publicly available content. The nonexpressive copying that occurs in the course of training is permitted in the EU and the U.S. and, to a less clear extent, in the UK. The TDM exemption the Government is considering would bring the UK into line with those other countries and provide much needed legal certainty for AI companies based and wanting to operate in the UK.

Requiring an exhaustive process to determine whether material is protected by copyright and where the copyright sits before using it in the AI training process would make the development of some of the most important, foundational models impossible. With development not taking place, important innovation that could support technological and economic progress would be frustrated and there would be no revenue opportunity for rights holders.

What can policymakers do to give creators more control over copyrighted works?

To the extent that policymakers are concerned that copyright holders are not being given the ability to authorise and prohibit the use of their works, the important mechanism is the ability to “reserve” their “rights” or “opt out” (as exists in the EU).

Leading AI developers already enable companies to opt out of text and data mining without opting out for other purposes such as search engine indexing. While it will be important to get the details right, based on a standard that has wide market-driven uptake and is machine readable at the source, a requirement for developers to respect an opt out—which builds on these sector initiatives—might be the best opportunity to maximize innovation in AI while reassuring rights holders that they will be able to control the use of their content.

Is there an information asymmetry between AI developers and rights holders?

AI developers will better understand how their models work, while rights holders will better understand their content and its context better. By the nature of a research exercise such as AI development, and the complexity of AI models, neither side will have a perfect understanding of how a given dataset can be used. However, there is no reason to think that AI developers have a general or structural advantage over rights holders.

Would “transparency” over AI model inputs improve outcomes?

The specific type, amount and weight of content used to train AI models can make a difference between the performance of one model or another. In such a highly dynamic market, publicly sharing the exact content used to train a AI model is very sensitive information that requires strict protection. There are a range of risks with any transparency requirement, growing more severe as the requirement demands more granular information, including:

- Security: if the functionality of AI models is made public then this will create opportunities for people to exploit those models and how they are trained. To the extent AI models are used to support important services over time, this could then have important consumer or security impacts.

- Competition: if developers cannot retain commercial secrets about how their models are developed, it will be harder for innovative developers to compete and stay in the market by differentiating their approach.
- Cost: sharing information about the very broad sets of data used to train the most significant models will be a major and expensive undertaking and undermine the dynamism of the market, in which a large number of models are currently being developed by a diverse range of companies.

Mitigating some of the risks could magnify the compliance costs and could prove impossible. Failure to protect AI providers' trade secrets and business confidential information in any broadly shared disclosure will confront companies with a difficult decision: either share sensitive business information and enter the UK market or avoid the UK market altogether. Given that large jurisdictions such as the U.S. and Japan have thus far not taken this approach, avoiding the UK market is likely to be the choice made, particularly for the leading models that the Government is keen to ensure can be developed here.

This could have further impacts on the quality of AI services used by UK consumers and businesses. Research exploring the impact of overlapping AI and copyright laws in the EU has found that "to "avoid licensing, it may be economically attractive for developers to train their algorithms on older, less accurate, biased data, or import AI models already trained on unverifiable data."[1]

It is also unclear what purpose such a requirement would address:

- Under this framework with a new TDM exception, companies should be transparent about their opt-out policies. As web publishers have control over whether their content can be included in training datasets, it makes little sense to have additional transparency requirements linked to who should receive remuneration.
- If the intention is to police opt outs, it is unclear why AI developers would comply with extremely challenging transparency requirements but not with relatively simple opt outs.
- If the intention is to police outputs that are copyright infringing, those outputs themselves (i.e. the video, audio or text outputs that users get from a model) would provide the evidence needed for any legal action.

Transparency requirements, particularly if overly granular and draconian, undermine the purpose of a TDM and other measures to attract leading AI development, without serving any real purpose for rights holders.

[1] Kretschmer, M. & Margoni, T. (2022) A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology, *GRUR International Journal of European and International IP Law*, https://academic.oup.com/grurint/article/71/8/685/6650009#368301263

Is a non-commercial exception sufficient?

The UK has a non-commercial TDM exception but this is not sufficient to meet Ministers' aspirations in this sector:

- Leading models are being developed commercially and the Government will not realise its ambition for the UK to be a jurisdiction in which those models are trained if it is restricted to non-commercial development. This will have implications for UK innovation and growth.
- The Government is keen to encourage the commercialisation of work done by the UK's world-leading research institutions and sees them as a source of economic growth. This will not be able to take place if regulatory requirements mean that models have to remain non-commercial.
- Many non-commercial and science and research organisations are partnering with commercial entities given the diverse capabilities needed for important projects. Unnecessary legal distinctions between commercial and non-commercial entities will undermine these partnerships.

How does this affect other public policy priorities?

There are two main Government objectives affected by this regulatory debate:

- Economic growth: if the UK is not an environment in which it is practical to develop leading AI models from a regulatory perspective then (a) other measures to promote economic growth will be less effective, given the broad technological opportunity AI represents; and (b) more specifically, other measures to encourage AI development in the UK (addressing important requirements such as compute) will be less effective.
- Public services improvements: the AI development that is likely to suffer the most is development to address UK-specific needs such as the needs of UK public services.