Copyright, Data and AI – What are we trying to achieve anyway?

The past year has seen a whirlwind of debate, consultations, guidelines, roundtables, and listening sessions on copyright. It's even been the topic of discussion in the pub – well among my IP friends at least: 2023, the year copyright became interesting. In the midst of this activity, it's important that we don't lose sight of what we want to achieve … which is what, exactly?

It is understandable that every industry is thinking about what impact AI will have on their business, as businesses consider what new roles may be required, and what operational changes could be made. There is clear excitement and interest about the potential new opportunities AI presents.

Questions concerning copyright and AI training seem focused on trying to pin down the specific technical processes involved in data analysis, in order to definitively assess if an act of copying has occurred. We have only to look at how machine learning methods have developed, and continue to develop, to know that attempts at this form of analysis could be out of date before the ink is dry, or at least before the next training epoch is complete. A more sensible approach to assessing copyright issues would be to determine if a copyright infringement is actually generated as an output.

In the context of training AI models on internet data, the issue is, according to some, about figuring out who gets paid – but is it? Even if compensation for training AI was required, what could that look like? Akin to having to license every linked page on the internet, every page touched by a search engine – what licence fee would be appropriate for a copyrighted work in context of the vast corpus of the internet? What is the cost of licensing a single token, in amongst the gazillion or so tokens used to train a foundation model? If licensing had been required for the internet, the internet as we know it today would never have existed. Even if AI could survive a licensing model for training, licensing models would prohibit open source AI development, entrench data monopolies, and lock AI development up in the hands of a few.

The most worrying aspect of this debate is the failure to recognise the potential impact to innovation if legislation is introduced that prevents technical analysis of publicly available copyrighted works. AI is already fundamental to how we innovate, it is not new – we have been using AI to innovate for years, including training AI models on

copyrighted works. The scientific advances that generative AI is now unlocking are truly remarkable. Only last month we saw that the use of AI had helped find a new material that could reduce lithium use in batteries by 70%, narrowing 32 million candidates to just 18 in less than a week. The AI for Science team at Microsoft Research has heralded foundation models as providing the next paradigm of scientific discovery, taking us from having to analyse systems of insoluble complexity, to allowing us to now unlock solutions from a vast problem space, so that suddenly we are far more capable of solving the world's most challenging problems. AI is speeding up innovation at a time when we really need to speed things up. The clock is running out for addressing climate change. AI can really help solve the climate crisis, but to do so it needs to be trained on vast amounts of data that are contained in copyrighted works. Restricting permissionless training on publicly available data could kill development of critically important technologies.

AI is here, it has been for a while, and it will enable amazing discoveries. But it's not an artist, it's not an inventor. It's not even an enormous, very clever monkey (thank you Andrew Marr for my favourite, but pretty unhelpful, analogy). Can we maybe just ditch the nonsensical narratives and see it for what it is? It's a tool, like any other, and we will find ways to create amazing things with it. The value in art is in human connection, and society will continue to optimise for this as we develop even more new ways of creating. Legitimate IP concerns, which relate to the potential for infringing *outputs*, are addressable, and rightly where we should be focused.

We need to ensure that this remarkable technology can be developed effectively and responsibly, and that the opportunity and benefits of AI are available to everyone. To achieve this, AI models will need to be developed with access to data, and people using AI will need access to data. Let's not stifle this opportunity at such a critical time. If we suffocate this technology, we will all miss out. Let's work together to figure out what it is we want to achieve, and then try to find the solution.

Sonia Cooper, Chair of IP Federation Data & Copyright Committee
13 February 2024