Roxburgh Milkins

Regulating AI:
How licence agreements can benefit AI developers and creators

Since the launch of ChatGPT in November 2022, the regulation of AI has become a hot topic.

People marvelled at the speed and ease with which ChatGPT gave answers to weird and wonderful questions and demands. But questions around regulation also quickly began to emerge.

One of the main complaints against large language models (LLMs) like ChatGPT is the way in which the models are trained. The software scrapes the internet for information and data to teach it the substance and style to deliver its answers.

The problem with that is that it raises questions of copyright infringement. The software is taking original content without giving credit to the creator, and benefitting from their work without giving the creator any reward.

In this blog, we look at a few of the problems raised by LLMs, and how website owners and AI operators alike can protect themselves against those risks.

Breach of terms

You may have seen recently that the popular parenting website Mumsnet has launched legal action against Open AI, the developer of ChatGPT.

Their complaint is that OpenAI scraped content from its website to help create ChatGPT, and teach it how to answer questions in a conversational tone.

OpenAI did not ask permission to scrape the site, and in doing so it breached the terms of use of Mumsnet's website, and infringed IP rights.

Mumsnet's issue is that this behaviour threatens their business model (paid for advertising) and existence. If users can get the answers from a chatbot, they don't need to visit the original site. Advertising revenue falls, etc. etc.

The way around this problem would be to agree a licensing deal to use the data. That way, the original creator is recognised and remunerated, and the AI developer has permission to use the material.

Copyright infringement

In a recent submission to a House of Lords inquiry, OpenAI recognised that it would be "impossible" to train today's leading AI models without using copyrighted materials.

The AI models are using copyrighted materials, often without permission, to generate their own content and present it as original. The original creator gets no recognition or financial reward, for the unsanctioned use of their original ideas, designs, written work etc.

This is a widespread infringement of copyright. Again, effective licensing would be one way to level up the playing field.

Gender bias

While it's not a legal point, Mumsnet raise an interesting point that the majority of content on the internet is written by men. That means that the data scraped by developers is training AI in a more male-oriented point of view. As Mumsnet puts it "AI models have misogyny baked in."

As AI develops, it will have to consider whether it can address any inherent gender bias that it has learned through fairly indiscriminate use of content from the internet. Mumsnet has said that they are keen to work with AI to address gender bias in coding, provided that use of their content is subject to a suitable licensing deal.

The framework for regulation

All of this leads to the question of regulation. The European Parliament formally adopted the EU Artificial Intelligence Act in March 2024. This Act takes quite a conservative approach to safeguarding. In the UK, under the previous Conservative Government, a more laissez faire approach was taken with a White Paper that advocated a "pro-innovation" approach.

However, the UK has not yet managed to address the problem of copyright infringement or breaching terms of use. A House of Commons Special Committee report published in January 2024 noted that the reproduction of copyright-protected works by AI will infringe copyright, unless permitted under licence or an exception.

There is support for a licensing regime. The report recommended that the Government "proactively supports small AI developers in particular, who may find difficulties in acquiring licences, by reviewing how licensing schemes can be introduced for technical material and how mutually-beneficial arrangements can be struck with rights management organisations and creative industries trade bodies."

Those recommendations still stand now that we have a new Government. The Labour Government has proposed to focus regulation on companies responsible for the most powerful AI models. We might expect a statutory code to come into force under this Government, which would require AI companies to share testing data with Government.

Under that code, companies would have to supply the AI Authority with third-party data and intellectual property used in the training of their models and ensure that any such data and IP is used with consent from the original source.

What it means for websites and creators

The best course of action for websites and creators at the moment is to have robust terms of use in place.

There should be a clause that says something along the lines of "no part of this site may be distributed, scraped, or copied for any purpose without the express approval of the website owner."

Once regulation, or a statutory code is in place, there may be further actions that you can take to protect your intellectual property.

There are also some technical solutions available. Open AI has enabled website owners to block its GPTBot crawler, although this does not apply retrospectively to content which has already been scraped.

Advice to AI developers

Of course, AI developers want to train their systems using quality data. But if you're an AI operator, you should be aware that scraping may breach the terms of use of the original owner.

To protect yourself from legal action you can request licences and permission to use the material.

If you'd like advice on any of these issues, please contact us at **info@roxburghmilkins.com**.

Call Ian Grimley on 0117 928 1915 or 07788 584308
Email: Ian.Grimley@roxburghmilkins.com
Visit: www.roxburghmilkins.com