WIPO Conversations on AI
Dr. Ana Ramalho, Copyright Counsel (Google)

AI is at the core of what we do at Google - we have been focusing on AI since its inception, and we are aware of the potential of AI to help people at scale. For example, AlphaFold, which covers 200 million predictions of protein structures, has helped with cutting plastic pollution, or tackling antibiotic resistance.

But today I'd like to focus on the backend, on how models, and in particular large language models, work. Models 'learn' in the same way that humans learn to read, write and acquire skills through reading. To train a model, a developer will input training data to the algorithm. The algorithm will then represent features of that data by assigning weights to said features, but the data itself does not 'stay' in the model, as the model does not store copies. Models follow by either categorizing or predicting what comes next, but there's no copying as such. Partly because of this complex workings, training takes a long time, is quite costly, and is hardly irreversible.

When a user queries a generative AI model, the answer will be the most probable, accurate answer from a statistical perspective, based on what the model has learned. That is why the importance of training is in the bulk, not in a particular piece of data - the model learns from quantity, so as to understand what is statistically more probable to be the correct answer to a particular question posed by its user.

This technical process has several implications for copyright law, but one important one is that there is no storage or collage or even tokenization of the training data. Models do not contain copies of individual elements of training data, they contain weights and functions. There's merely a representation of, or a learning from, the training data. Many legislations around the world have taken this into consideration, and exempted machine learning, or text and data mining, from the exclusive rights of copyright owners.

But there's another reason for this cautious approach to machine learning by some regulators and policy-makers. Yes, GenAI can function as your creative assistant and give you ideas for your child's birthday party, or help you organize your thoughts if you want to write a novel. But beyond that creative skill, AI can also - for instance - help healthcare workers deliver quality care at scale, by unifying data across various sources (including research papers), making it easier to search and analyze the most relevant results. So, not only do models not contain their data for purposes of copyright law, they help make the world a better place by having impact at scale.