WIPO Conversations on AI - Tenth Session (2024)
Dr. Ana Ramalho, Senior Copyright Counsel (Google)

Good afternoon everyone, and thank you for the opportunity to be here. I am Ana Ramalho and I am a Senior Copyright Counsel at Google. I'd like to focus my intervention on what AI-generated outputs are not: and that is, copies. AI outputs are not copies of existing works. This becomes evident once one peers into the inside of a model, which I propose to do in the few minutes I have. Let's walk through the process that will lead to the production of a text output as an example.

The model is not a database of works, nor is it a technology that carries out bitwise copies of the data which is used to train it. So what does a model do with the training data? It assesses the proximity, frequency, and other elements of data (or portions of data) called tokens, and the structure of text. It does this so it can establish probabilities of relationships between the tokens. During the training process, the model will adjust its own parameters (numerical weights and math functions) to reflect these (mathematical) relationships. Once the model has adjusted its parameters to accurately reflect these relationships, it can then be used to generate new outputs based on those parameters. The model itself is thus a large network of weights, of numerical values, that represent these learned relationships; in other words, a trained model has learned to identify relationships and patterns among words.

When this learned model is prompted to generate output, it will produce a response which is the most statistically likely to satisfactorily address the prompt that the user has entered, based on what it has learned during the training process. Importantly, what the model has learned is not the substantive content itself of a particular piece of training data, but rather the mentioned relationships between the tokens from many (many) elements of training data. Meaning, the model does not learn from any piece of individual content, but from all the content on which it is trained. It follows that, as a rule, no particular, individual work is essential for the training of the model - the value lies in the total collection of content, works, or datapoints needed to train an AI model.

Following a prompt provided by the user, a trained model will then be able to produce new content due to a combination of its internal mathematical function and information about the data (again, not about the substantive content of the data, but information on patterns, or for instance relationships between words, etc).

So in a nutshell, there are 2 fundamental points here: (1) there are no copies of training data in the model, and (2) any output is generated, and not a per se copy of the training data. Connected to these points, and to the tremendous amount of data used to generate LLMs, an output produced by an AI model cannot reliably be traced to a specific piece of training data - although some research groups are currently dedicated to this issue, this particular field is still in its infancy (and we would therefore caution against regulation on such a volatile area).