**WIPO Conversations on AI - Ninth Session**
**Dr. Ana Ramalho, Copyright Counsel (Google)**

In the previous WIPO Conversation, I talked about how models work, and in particular how they do not store their training data. Today I would like to follow-up to talk about a related technical issue, memorization, and why memorization is not reproduction or a copyright-relevant act.

The term "memorization" is used by ML researchers to express a likelihood that outputs from a model might be similar to elements of the model's training data. The ML community uses that term in that testing context, but there is little in the published literature to make this point clear - that is, it is rarely said that the technical definition of memorization is different from the colloquial one.

Some technical papers have argued that in certain conditions models may memorize their training data. But again, it is important to understand what ‘memorization’ means in this context. As mentioned, the definition of memorization in the technical context of trained models refers to the likelihood that, by designing an attack on those models, they can be made to generate an output that looks almost exactly like training examples. Or, to put it differently, technical memorization looks at whether a model can be prompted to generate a piece of content that it was trained with.

The answers to these questions are more nuanced than one might think. Where reconstructing a piece of training data was found to be possible, several conditions needed to be met: first, the person prompting the model to produce that content had to know or to at least suspect that the specific content was part of the training data to begin with. Second, the model had to be prompted with specific appropriate instructions, which is the so-called ‘attack on the model’ that I just referred to. Third, the success rate will depend on how often the specific piece of content appears in the dataset - since models are probabilistic machines, the more often a piece of content shows in the dataset, the more weight the training process will attribute to the characteristics of that content, and hence the more likely it is that during generation the model will assume that the content is statistically relevant to the response to a prompt. More recent studies have shown that memorization cases predominantly occur when either the training set is small, or when there are a number of duplicate training samples - further reinforcing the fact that the output of Generative AI amounts to the most probable response to a prompt, statistically speaking. Indeed, these recent studies have found that it is currently unknown whether memorization would take place on larger and more diverse datasets.

So, in a nutshell, ‘storing’ training data and ‘memorizing’ training data are two very different notions, conceptually and in practice. Models are unable to do the former. Models might be able to do the latter, in very particular conditions that are usually not found in the real world, and certainly less so in relation to Large Language Models. It is important not to mix the colloquial definition of memorization (which calls to mind the idea of a copy) with its technical definition (which does not imply a copy or a storage of that copy in the copyright sense). Model “memorization” is not an act of reproduction, it is the measure of whether the model can generate output that is similar to its training data.