Digital Fools' Gold – The Rise (and Potential Fall) of Rightsholders' Efforts to Cash In on the GenAI Gold Rush

September 11, 2024 Lee Johnston

PRACTICES AI and Deep Learning

The dominance of generative artificial intelligence ("GenAI") in today's popular culture is undeniable. Practically every 30 second advertising spot during the Olympics touted the mind-blowing capabilities of yet another GenAI-assisted tool promising to take us to new levels of efficiency and creativity. Business journals report that massive amounts of capital investment are being directed not only to companies who create GenAI-assisted tools, but also to the companies like Nvidia, whose products support the development of these AI-assisted tools.

But what about compensating the individuals who provided the fuel—the massive amounts of data needed to train AI models—without which the "AI revolution" would have never happened? Recent court decisions suggest that the answer to this question will turn on whether the harvesting and use of this training data constitutes "fair use" under the Copyright Act.

Brief Overview of GenAI

"Generative AI" systems, such as the Generative Pretrained Transformer (GPT) and Large Language Model Meta AI (LLaMA) language models and the Stable Diffusion and Midjourney text-to-image models, were built by ingesting massive quantities of text and images from the internet. Today's GenAI models are machine learning models trained on social media posts, books, articles, photos, digital art, music, software, and more. Rather than simply classifying these diverse inputs and generating metadata about them—as previous generations of machine learning systems have —GenAI models can produce new digital artifacts: new text, new art, new music, and new software.

Generally speaking, the GenAI-assisted tools are developed and work to create prompt-generated outputs through two transition processes. The first transition is the *training* process. In training, the system is exposed to a large amount of data relevant for its purpose, known as the *training set*. In many instances, this data is harvested from publicly available Internet websites, including social media sites like LinkedIn, Facebook, and Instagram, through an automated process known as data scraping. Thus far, efforts to legally enjoin data scraping of this publicly available (*i.e*, non-password protected) data via contract (e.g., breaches of website Terms of Service) or other non-copyright legal theories (e.g., unfair competition, unjust enrichment) have been rejected by courts.

Read the full article here.