HarperCollins Publishers, LLC
William S. Adams

Artificial Intelligence and Copyright

Notice of Inquiry and Request for Comments

(Docket No. 2023-6)

HarperCollins Publishers welcomes the opportunity to provide these Reply Comments to the Notice of Inquiry and Request for Comments on the intersection of copyright law and artificial intelligence. HarperCollins is not only the oldest trade publisher in the United States but is also keenly focused on the latest developments in technology and copyright law. Our comments below are fully supportive of and consistent with those contained in the Submission of the American Association of Publishers and reiterate its conclusions:

1. That copying and use of copyrighted material for the training of Gen AI requires consent;
2. That unauthorized use of copyrighted material runs counter to the policy of copyright law;
3. That Gen AI companies must be transparent about the specific works of authorship used by them in training and tuning their models; and
4. That a legislative solution may be essential if courts fail to protect copyright owners from widespread and unauthorized use of their works by Gen AI companies.

Our brief comments below are in opposition to much of the post-hoc self-justification expressed by the Gen AI companies that seek to profit off the expressive content of authors.

Imbedded in the Constitution, copyright law has at its foundation the objective of promoting the development of the useful arts by granting a property right to authors and others who invest in creating them. It balances the interests of the creators with those of later artists and writers, encouraging and protecting the former while allowing the accretion of new works that may in certain circumstances make limited use of the prior work.

Gen AI companies threaten to upend this balance by reading or ingesting virtually all books ever written, then producing a torrent of computer-generated material at previously unimagined speed and volume. Nevertheless, in their submissions, proponents of Gen AI claim that their AI models do not rely on the expressive content of the thousands of authors' works they have used for training and finetuning their models. Their submissions minimize the importance of the expressive content of the books they copy, creatively describing their use in other ways. For some, it is akin to mining: "extract[ing] facts and statistical patterns" (Andreessen Horowitz at 6), or "extracting unprotectable elements" (Anthropic at 7), or "extracting unprotected information *about* the English language – i.e., the correlations, patterns, and relationships among the 26 letters of the alphabet and the 1 million English language words" (BSA at 8). As if these are minerals lying deep under the surface of some previously uncharted and unclaimed territory.

In other submissions they study the expressive works themselves, "analyz[ing] the structure and syntax of language . . . to discern the statistical relationships between words." (OpenAI at 11). Or they use the expressive texts to "generat[e] [vectors] to represent not just words but information about the semantic and contextual meaning of the words and their relationships to other words in the vocabulary." In other words, they use the underlying expressive texts to generate *new* expressive texts. (Microsoft at 6).

In fact, what all these examples are extracting or analyzing is the way in which authors have expressed themselves – precisely so the Gen AI companies can emulate that expression convincingly for AI users. Rather than create their own sentences, paragraphs and stories to train their systems or limit themselves to using public domain materials, the Gen AI companies have pirated the copyrighted work of others and now strain to recharacterize their cost-cutting short-cuts.

Moreover, the Gen AI companies not only are using authors' books for initial training but will remain reliant on the ongoing use of the underlying expression created by authors. Without the continuous training provided through the use of human-created works, Gen AI-generated writing that trains only on its own output rapidly deteriorates into "gibberish." Rahul Rao, AI-Generated Data Can Poison Future AI Models, *Scientific America*, July 28, 2023. Not only have the Gen AI companies ingested copyrighted works to create their models, but they must continue to rely on the copyrighted works of human authors who have not given their consent to prevent the Gen AI models from corrupting their own output with inferior Gen AI-generated content.

We appreciate the opportunity to provide our views and look forward to working with the Copyright Office on this matter.

HarperCollins Publishers, LLC

December 6, 2023