Raina Siladi

October 30, 2023

Question 8.3, Notice and Comment for Docket # 2023-6, U.S. Copyright Off.

Artificial Intelligence and Copyright

The fair use analysis applies to Artificial Intelligence (AI) models with “market-encroaching uses” that produce “expressive” work.[1] AI Policy writer Ben Sobel defines market-encroaching uses that “ingest copyrighted expression for a purpose that endangers the market for that very expression.”[2] Market-encroaching uses differ from overtly noncommercial or research uses in elements one and four of 17 U.S.C. § 107: (1) “the purpose and character of the use” and (4) “the effect of the use upon the potential market for or value of the copyrighted work.”[3] Nonexpressive or purely fact based output summarizes or indexes copyrighted input, generating output “about”[4] the original. Nonexpressive use has “historically been deemed fair use.”[5] However, when copyrighted materials inputted into an AI model generate “novel and expressive” output (element (1)), the output can be licensed more cheaply than the original author’s work (element 4). The AI’s purpose can be considered non-transformative, and market-encroaching under element (1), if the output is designed to be sold to the same buyers interested in the inputted copyrighted works. For example, Sobel suggests AI could “even force human composers out of some segments of the music market”, like “stock” and background music.”[6]

Sy Damle, a prominent venture capitalist, has a different perspective.[7] He insists that “it would extend copyright law beyond its recognizable bounds to say that creating something in the style of an author” is infringement.[8] His example contrasts with Sobel’s music premise: “if I were to compile the works of Stephen King simply so I can emulate his style of writing, I don’t think anybody would say that I have infringed.” And so if a computer does the same thing, no infringement occurs.[9]

Derek Slater, founder of American nonprofit Creative Commons, advocates for common ownership of information. He describes a distinguishing factor in the commercial use analysis. The makers and sellers of the AI tool are generally not the users. Who infringes? “We have tools like secondary liability to think about whether the tool creator themselves is contributing or not. I think, in most cases, they aren't. It's the user who's doing it.”[10] The legal framework for a human infringer cannot apply identically to an AI model. The AI tool is a unique creature, separate from the training dataset and the third party human user. Each AI model has a different algorithm. Computer coders have styles “unique as fingerprints.”[11] Perhaps the programmers’ transformative role applies to debates about simple input and output.

The AI “systems ingest massive amounts” of an enormous training dataset without identifying or sourcing the copied works.[12] Google Books “contains twenty-five million machine-readable copies of print books.”[13] “In 2005, a Google employee . . . explained, ‘We are not scanning all those books to be read by people . . . We are scanning them to be read by an AI.’”[14] Before corporations “invented big data as an idea,” public archives had begun digitization. [15] Archivists could not archive, appraise, and select material without “machine reasoning.”[16] AIs prompted to compose material in the style of a particular author are trained on these massive inputs, not only on the particular author’s work.[17] The AI tool trained in this manner moves even farther afield than Damle’s Stephen King example. AI can absorb and process tremendously more than human intelligence.

Question 8.3 asks how fair use applies to commercial AI models trained on noncommercial or common datasets. This question may no longer apply in the near future. According to MIT’s Sloan Center, “The more [training] data, the better the program”[18] and, “machine learning is best suited for situations with lots of data — thousands or millions of examples.”[19] The model needs to correct itself with additional, updated data. Chatbots, for example, need to learn "from records of past conversations to come up with appropriate responses."[20] Confining AI to noncommercial or research data or to any restricted dataset will limit sophistication. AI need access to as much data as possible.

Corporations often do not report the sources of their training datasets. The necessary size of the datasets combined with the "lack of transparency to users and to content owners"[21] complicate the fair use analysis. "So many of these systems may fail to provide source attribution, so when responses are given to users," the original source is unknown.[22] Original source may not have the same meaning for AI. Models train to make choices. The simplest choice is yes or no. AIs have to train on data satisfying a range of choices. The content of a particular question and its specific output requires AI reference to data both inside the content box *and* outside of it. AI is a type of intelligence, or processing. It does not simply categorize and copy.[23]

Corporate AI owners purposely create nonprofit foundations to launder data. The AI trains on data from the third party research foundation, and the original copyrighted data is made clean by the nonprofit.[24] In the *Fair Use Index*, the U.S. Copyright Office writes about element (1), purpose or character of the use: "Courts look at how the party claiming fair use is using the copyrighted work and are more likely to find that nonprofit and educational and noncommerical uses are fair."[25] Training datasets consist of material that may appear in the output response to a particular query, but also serve to teach the model what should not be included in said output response. This makes the calculation of royalties difficult. Code that identifies every chain of text plagiarized from a copyrighted work could help construct a system to compensate original artists. Corporations should not be able to launder data to avoid paying copyright royalties.

However, the same problem arises in view of the size of data that the most sophisticated AI train on. Market-encroachment appears less sharply defined. A standard that examines output in comparison to a human creator's original work is advisable. However, data laundering that allows corporations to obscure the AI output's original source in the output itself may not be covered by this standard.

Ultimately, courts will decide when the corporations that own AI models reimburse human creators. No doubt money will change hands and most of that money will remain with those that already have more. Sobel's legal analysis makes sense. If courts choose the market-encroaching standard, individual authors will win damages. The damages calculation will be more complicated, or perhaps impossible, the larger the dataset. Some rare situations will arise where the plagiarism is obvious enough to compensate the original creator. In the future, all content creators may write boilerplate and collect royalties. Meanwhile, AI will take over more markets and products. A trend may "move towards closing down access" to data for more than just copyright infringement excluded by fair use.[26] "Not only will the open-source crowd be cut adrift—but the next generation of AI breakthroughs will be entirely back in the hands of the biggest, richest AI labs in the world."[27]

Future AI will use "AI to make AI" leading to "artificial general intelligence—machines that can outthink humans."[28] AI may create "technical solutions humans would not think of by themselves, inventing new and more efficient types of algorithms or neural networks—or even ditching neural networks."[29] Finding the trail back from an AI creation to a human author may soon become impossible. Hopefully we will not lose our humanity as well.

[1] Benjamin W. Sobel, Comments Letter on Intellectual Property Protection for Artificial Intelligence Innovation PTO-C-2019-0038 (Dec. 15, 2019), https://www.bensobel.org/files/misc/Sobel-AI-USPTO-Comment-PTO-C-2019-0038.pdf.

[2] *Id.* at 2.

[3] *Id.* at 2 (quoting 17 U.S.C. § 107)

[4] Elise De Geyter, *Inside Views: The Dilemma of Fair Use and Expressive Machine Learning: An Interview with Ben Sobel*, INTELLECTUAL PROPERTY WATCH (Aug. 23, 2017), https://www.ip-watch.org/2017/08/23/dilemma-fair-use-expressive-machine-learning-interview-ben-sobel/ .

[5] *Id.*

[6] Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM.J.L. & ARTS 45, 79 (2017), https://www.bensobel.org/files/articles/41.1_Sobel-FINAL.pdf .

[7] Sy Damle, Transcript of Proceedings, Copyright and Artificial Intelligence Literary Works, Including Software, Listening Session, LIBRARY OF CONGRESS, UNITED STATES COPYRIGHT OFFICE 117 (April 19, 2023).

[8] *Id.*

[9] *Id.*

[10] Derek Slater, Transcript of Proceedings, Copyright and Artificial Intelligence Literary Works, Including Software, Listening Session, LIBRARY OF CONGRESS, UNITED STATES COPYRIGHT OFFICE 115-16 (April 19, 2023).

[11] Phil Johnson, *CSI Computer Science: Your Coding Style can Give You Away*, Computerworld United States (Jan. 28, 2015 3:45 AM PST), https://www.computerworld.com/article/2876179/csi-computer-science-your-coding-style-can-give-you-away.html .

[12] Cynthia Arato, Transcript of Proceedings, Copyright and Artificial Intelligence Literary Works, Including Software, Listening Session, LIBRARY OF CONGRESS, UNITED STATES COPYRIGHT OFFICE 112-13 (April 19, 2023).

[13] Sobel, *supra* note 5, at 71.

[14] *Id.*

[15] Colavizza, *Archives and AI: An Overview of Current Debates and Future Perspectives*, 15 ASSOC. FOR COMPUTING MACHINERY J. ON COMPUTING AND CULTURAL HERITAGE, no. 4 at 1-2, Dec. 2021.

[16] *Id*. at 2.

[17] Arato, *supra* note 11.

[18] Sara Brown, *Machine Learning, Explained*, MIT SLOAN SCHOOL IDEALS MADE TO MATTER, ARTIFICIAL INTELLIGENCE (Apr. 21, 2021), https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained#What%20Is%20Machine%20Learning? .

[19] *Id*.

[20] *Id*.

[21] Cynthia Arato, Transcript of Proceedings, Copyright and Artificial Intelligence Literary Works, Including Software, Listening Session, LIBRARY OF CONGRESS, UNITED STATES COPYRIGHT OFFICE 112-13 (April 19, 2023).

[22] *Id*.

[23] Brown, *supra* note 17.

[24] Andy Baio, *AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability*, WAXY (posted Sept. 30, 2022) https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/ (cited by Nieman Lab, *What We're Reading*, Nieman Foundation at Harvard University https://www.niemanlab.org/reading/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/ (last visited Oct. 29, 2023)).

[25] U.S.COPYRIGHT OFFICE, U.S. COPYRIGHT OFFICE FAIR USE INDEX (updated Feb. 2023) https://www.copyright.gov/fair-use/ (last viewed Oct. 29, 2023 10:50 PM).

[26] Will Douglas Heaven, *The open-source AI boom is built on Big Tech's handouts. How long will it last?,* MIT TECH. REV. (May 12, 2023), https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta/ .

[27] *Id.*

[28] Will Douglas Heaven, *AI is Learning How to Create Itself,* MIT TECH. REV. (May 27, 2021), https://www.thesentientrobot.com/ai-is-learning-how-to-create-itself-by-will-douglas-heaven/ .

[29] *Id.*