



SCHOOL OF INFORMATION
102 SOUTH HALL #4600
BERKELEY, CALIFORNIA 94720-4600

October 9, 2023

Dear colleagues:

As the Copyright Office solicits feedback on the role of copyrighted material in both generative and non-generative AI, I want to offer comments of my own on the latter question in particular; I am an associate professor in the School of Information at UC Berkeley whose research focuses on designing empirical and computational methods for the study of culture (including fiction, music, film and television). I have given feedback to the Copyright Office in the past on the eighth triennial rulemaking on exemptions to the prohibition against technological protection measures under the DMCA.

In the course of my work, I use copyrighted materials to train **non-generative AI**. These materials are all lawfully acquired: I purchase books and scan them to create digital versions that computational systems can interact with; I purchase DVDs and break digital rights management on them following the protocol outlined in §37 CFR 201.40(b)(4).¹ I do so as a good faith actor, respecting the material rights of authors and other content creators, but also balancing the public good that comes from analyzing their collective works. The non-generative AI that I train on these materials cannot be used to reproduce the originals; these are systems that *measure*.

For works of fiction, one such system, called BookNLP,² finds the spans of text in a novel that refer to people and places (e.g., in *Call me Ishmael*, automatically identifying “Ishmael” as a character). This system has been used to identify that men as characters in books get three times more screentime than women over the two centuries between 1800-2010 (Underwood et al. 2018); it has shown that women are depicted in fiction in domestic spaces far more often than men (Sandeep et al. 2023) and women as characters are also represented as the linchpins of information flow (Sims et al. 2020). Others have used it to uncover the heteronormativity of relationships in books (Kraicer and Piper 2019). In all of these cases, we have new knowledge about the world, and its representation in fiction, partly as a result of a system that has been trained on copyrighted texts.

For film and television, these non-generative AI systems identify the guns that show up on screen, along with the actors. The input to these systems is a sequence of video frames, and the output is a sequence of coordinates where the guns and actors can be found. While this research is ongoing, others have used similar systems to measure disparities in screentime between men and women as actors,³ including disparities in speaking time (Guha et al. 2015).

One might ask whether we can train non-generative AI on public domain materials to carry out these measurements. We cannot. At the core of this kind of research is *measurement validity*—the ability of a measuring instrument (here, an algorithm) to accurately measure what it is designed to measure. At the time of

¹[https://www.ecfr.gov/current/title-37/part-201#p-201.40\(b\)\(4\)](https://www.ecfr.gov/current/title-37/part-201#p-201.40(b)(4))

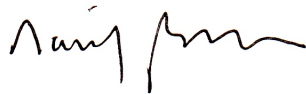
²<https://github.com/booknlp/booknlp>.

³https://about.google/intl/ALL_us/main/gender-equality-films/

writing, public domain in the United States generally applies to works published before January 1, 1928. For the foreseeable future, the most recent public domain materials will always be rooted 95 years in the past—a world that is vastly different from today, not only in terms of what gets depicted (such as the roles and occupations of men and women) but also in terms of who gets to do the depicting. (The History of Black Writers project⁴ alone identifies roughly as many novels published by Black authors over the period 1800–1930 as in the period 1984-1986.) If we are to analyze works published in 2023 correctly in order to be able to draw knowledge from that measurement, we cannot do so with a system that encodes a view of the world that ends in 1928.

This brings me to my fundamental point: the research that I am describing treats works in copyright as our *object of study*. We need to understand the novels and films that are being created right now: we measure contemporary novels in order to tell us something about the world in which those novels are created, and the disparities that exist within it; we measure film and television to likewise reflect back on how representation on screen matters for shaping the attitudes and beliefs of a contemporary audience who watches it. When copyrighted materials are our object of study, we must be able to train non-generative AI on those materials in order to study them. Removing the ability to train these kinds of algorithmic measuring devices on copyrighted material would effectively remove our ability to study them well. The concept of fair use balances the rights of content creators with the public good, and I count this kind of work—of shedding light on the disparities that exist within contemporary culture—as exactly that kind of good.

Sincerely,



David Bamman
Associate Professor
School of Information
University of California, Berkeley

⁴<https://hbw.ku.edu/novel-collections>