Library of Congress Public Comments
Artificial Intelligence and Copyright

https://www.regulations.gov/commenton/COLC-2023-0006-0001

- 1. COLLECTION AND CURATION OF DATASETS The use of copyrighted works to train AI Models in either generative or non-generative training sets without a license is a copyright infringement. This is flagrant outright theft, because Big Tech knows what copyright law is but choose to ignore it. In generative AI, the crawlers scan your data and throw it away retaining knowledge in their database. No permission access is requested by the crawlers although they should be required to contact the content publisher. Our government has evolved laws based off copyright and property rights which BigTech chooses to ignore. If the AI crawlers do not have a license to train on copyrighted content, they have no right to crawl the content. As a consequence of Big Tech's AI unauthorized internet scraping, millions of creators must publish their work for fear of ingestion by a generative AI web-crawler. Freedom of speech and publication are now foregone rights unless you are willing to give your work away. Without creativity our society dies. DMCA codified this behavior with social media and it destroyed publishers.
  - An automatic license at a legislated price is not acceptable; creators should have a right to negotiate their own license fees. All data is not equally valued.
  - I am opposed to the idea there should be a universal license fee to allow AI web crawlers to ingest copyrighted content and pay a universal fee that is arbitrarily assigned. All data is not equally valued and the copyright holder should be able to establish their own licenses.
  - Machine-to-machine licenses can solve this problem and the vocabulary and software expression models to implement frictionless commerce have already been developed by non-profits in the W3C POE and ODRL working groups. We should

move to adopt machine-to-machine licensing and use technical identifiers to validate copyright holders, assignees and licenses. If our technology companies are advanced enough to execute generative AI, then they are advanced enough to implement frictionless licensing to pay for the input materials in a manner consistent with law.

- Big Tech has steadfastly refused to honor copyright notices in content and have built billion dollar social media and search businesses off of free content. I have worked on many volunteer groups seeking to solve this problem, but Google and the big three steadfastly refuse to read or acknowledge copyright. Should we allow this to continue? NO. There has to be some legal basis for property rights in technological innovation.

- REMUNERATION

What kind of renumeration and how is it feasible? I do not believe that legislation should mandate the amount of renumeration for AI scraped data, especially when it relates to copyright-registered data.

The feasibility answer lies in technology. The Library of Congress should issue embeddable identifiers that enable assigned copyright holders or licensees content to be validated to their copyright registrations . After a copyright holder embeds the identifier in their work and publishes it, a user could read the identifier and validate the copyright registration belongs to the publisher or is represented by a license. The identifiers would point back to the public catalog of the Library of Congress. Ads who violated reading the registrations should be prohibited. Copyright Registration Assignments and transferred should be recorded and reflected on the public LOC catalog. There will be no excuse for saying you didn't know it was copyrighted. We should not rely upon browser extensions to see if something is legal. If a creator has

registered the copyright, then the Library of Congress must increase their tooling to validate the copyright registration to the copyright holder for any machine reading it.

- RECORD RETENTION

The brazen and appalling behavior of current generative AI applications is disheartening. Swept away by their own power, AI companies steal from published sources, then throw away the resources to hide their misdeeds. This is why all current generative **AI models should be taken offline now**. The datasets need to be rebuilt from the ground up in a manner consistent with copyright law. The idea that you can steal, destroy the records and walk away with the booty is the equivalent of financial fraud. As public scrutiny of AI source content increases, the foundational models become less transparent.I refer you to this recent Stanford University report on the lack of record transparency in current large language AI models shown in this link from the New York Times https://www.nytimes.com/2023/10/18/technology/how-ai-works-stanford.html and this report from Venture Beat. https://venturebeat.com/ai/how-transparent-are-ai-models-stanford-researchers-found-out/

- There should be an OPT-IN policy and **NOT** an OPT-OUT policy to permission AI crawlers whether generative or non-generative

2. SCOPE OF COPYRIGHT PROTECTION

Liability should be enforced against works generated using AI systems that pull from copyrighted work. Otherwise, a copyright registration means nothing. No copyright registration should be allowed for content scraped to a generative AI. Under the law, a willful copyright infringement carries a penalty of $150,000. Give generative AIs a choice: reveal all their source material and if it is found to contain copyrighted content, fine them $150,000 per infringement for willfulness. Alternatively, they can pull their datasets offline now and authenticate them properly.

- Outputs imitating a style of an Artist are akin to Deep Fakes, but will be very difficult to decide. I do not have any suggestions in this particular field. I do believe there are instances where a human's use of gnenerative AI could be altered enough that a copyright registration might be issued, especially if the creator's own authenticated work is used in the output seeking a copyright registration. This requires further discussion.

3. COPYRIGHT LIABILITY

In my opinion, all generative AI should taken offline now and rebuilt. There should be no compensation for the fact the AI has already been built. Theft is theft. To put the burden on the copyright owner, to track down and prosecute the AI is unfair.

If our government chooses to not force the current generative AI offline until they authenticate their sources, you shift the burden onto the copyright registrant or creator. This is unfair, particularly if the creator has followed the letter of the law and registered their copyright.

From my own personal experience my copyright-registered content was scraped from website that I did not place it on. The content was published on this website under a rights-managed contract I negotiated where the burden was on the licensee to protect the content. Here you have an example of a sophisticated big tech company who were supposed to abide by the terms of a written license agreement and instead they freely published it on the internet knowing it would be crawled. Three years later, I am still legally fighting this issue.

A potential unforeseen outcome if the US government decides that AI scraping can be done at set rate on copyrighted material, is that people who have previously negotiated licenses will allow the licensed content of the Creator to be monetized from the individual who bought the license.