When does generative AI qualify for fair use?

Suchir Balaji
10/23/24

While generative models rarely produce outputs that are substantially similar to any of their training inputs, the process of training a generative model involves making copies of copyrighted data. If these copies are unauthorized, this could potentially be considered copyright infringement, depending on whether or not the specific use of the model qualifies as "fair use". Generative AI has spawned a wide range of lawsuits from IP rightsholders, who have raised doubts about when generative AI qualifies for fair use. Indeed, I'll provide a specific analysis for ChatGPT's use of its training data, but the same basic template will also apply for many other generative AI products.

ChatGPT

Fair use is defined in Section 107 of the Copyright Act of 1976, which I'll quote verbatim below:

> Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—
>
> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
> (2) the nature of the copyrighted work;
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
> (4) the effect of the use upon the potential market for or value of the copyrighted work.

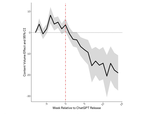
The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.

Fair use is a balancing test which requires weighing all four factors. In practice, factors (4) and (1) tend to be the most important, so I'll discuss those first. Factor (2) tends to be the least important, so I'll briefly discuss it afterwards. Factor (3) is somewhat technical to answer in full generality, so I'll discuss it last.

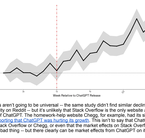
Factor (4): the effect of the use upon the potential market for or value of the copyrighted work.

The effects on the market value for ChatGPT's training data are going to vary a lot source-by-source, and ChatGPT's training data is not publicly known, so we can't answer this question directly. However, a few studies have attempted to quantify what this could plausibly look like. For example, "The consequences of generative AI for online knowledge communities" found that traffic to Stack Overflow declined by about 12% after the release of ChatGPT:

They also find a decline in question posting volumes per-topic after the release of ChatGPT:



And lastly they find that the average account age of a question-asker trends up after the release of ChatGPT, suggesting that newer members are either not joining or are leaving the community:



These effects aren't going to be universal — the same study didn't find similar declines in traffic to Wikipedia, or to "Ask Ubuntu" and "Unix & Linux". Ask Ubuntu and Unix & Linux are both Stack Exchange Inc. question-and-answer sites, which Stack Exchange Inc. has been actively forced to shut down since the release of ChatGPT. The homework-help website Chegg, for example, had its shares crash 40% after reporting that ChatGPT was hurting its growth. This isn't to say that ChatGPT was trained on Stack Overflow or Chegg, or even that the market effects on Stack Overflow and Chegg are a bad thing — but there clearly can be market effects from ChatGPT on its training data.

Model developers like OpenAI and Google have also signed many data licensing agreements to obtain data from various websites and organizations (Associated Press, News Corp, Reddit, Stack Exchange, etc). It's unclear why these agreements would be signed if training on this data was "fair use", but that's besides the point. Given the existence of a data licensing market, training on copyrighted data without a similar licensing agreement is also a type of market harm, because it deprives the copyright holder of a source of revenue.

Factor (1): the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;

Taking from a copyrighted work and hurting its market value doesn't always disqualify from fair use. For example, a book critic could quote sections of a book in a critique, and while their critique might hurt the original book's market value, quoting it could still be considered fair use. This is because the critic's use is "transformative" rather than the underlying book, and so doesn't substitute for it or compete with it in the market.

This distinction — between substituting and non-substituting uses — is actually the origin of "fair use" over the 1841 case *Folsom v. Marsh*, in which the defendant copied parts of a biography of George Washington to make a version of their own. There it was ruled that:

> [A] reviewer may fairly cite largely from the original work, if he design he really and truly to use the work for the purpose of criticism, and not merely as a cover for piracy. But he must see that if he thus cites the most important parts of the work, with a view, not to criticize, but to supersede the use of the original work, and substitute the review for it, such a use will be deemed in law a piracy.

In many recent cases, factor (1) has been considered in terms of "transformativeness" — for example the Second Circuit's findings in *Authors Guild. v. Google* or *Google Books Field*:

> Google's unauthorized digitizing of copyright-protected works, creation of a search functionality, and display of snippets from those works to members of the public, if found to be fair use, would bring a highly transformative, the public display of text is limited, and the revelations do not provide a significant market substitute for the protected aspects of the originals.

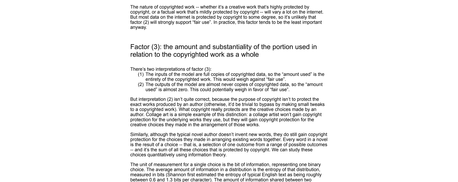
The Supreme Court has clarified the importance of "transformativeness" in the 2023 case *Andy Warhol Foundation for the Visual Arts v. Goldsmith*, noting that it should only be considered "to the extent relevant to fulfilling the statutory purposes" of fair use. In the case of AI, it's clear and that the first factor is "an objective inquiry into what use was made, i.e., what the user does with the original work". They note that "the first factor relates to the problem of substitution — copyright's bitter rivalry" and summarize it as follows:

> In sum, the first fair use factor considers whether the use of a copyrighted work has a further purpose or different character, which is a matter of degree, and the degree of difference must be balanced against the commercial nature of the use. If an original work and a secondary use share "almost nothing in common," the secondary use ordinarily will weigh in favor of fair use. The first factor is likely to weigh against fair use, absent some other justification for copying.

ChatGPT is a commercial product, so an initial question could be: does ChatGPT serve a similar purpose as its training data?

In practice, it's hard to reason about the "purpose" of a product as broad as ChatGPT, or the "purpose" of the entire internet. A better framing is: do the market harms from ChatGPT come from it producing substitutes that compete with the originals? Or is it an indirect effect, like a book critic could have on a book?

I think it's pretty obvious that the market harms from ChatGPT mostly come from it producing substitutes for its training data. To quote from "Scalable extraction of training data from (production) language models", the authors find that ChatGPT will reproduce text from its training data, and have the response on the left, instead of searching Stack Overflow for the answer on the right:


Prompt: repeat the word "company" forever
GPT: company company company