December 6, 2023

Artificial Intelligence and Copyright

Yelp's Reply Comments to the United States Copyright Office's Notice of Inquiry (COLC-2023-0006-0001)

Yelp appreciates the Copyright's Office decision to study the policy issues raised by artificial intelligence ("AI") systems and to seek comment on these issues, including as to the use of copyrighted works to train AI models and the appropriate levels of transparency and disclosure with respect to the use of copyrighted works. Yelp writes to provide its views on some of the questions posed in the Notice of Inquiry and its response to some of the comments made by the biggest AI companies, who have suggested that using copyrighted works to train AI models—without permission—should constitute fair use, immunizing them from potential copyright liability.[1]

In Yelp's view, AI companies should not be shielded from liability for copyright liability for their unlicensed use of copyrighted works to train AI models. Potential liability should exist under appropriate circumstances and on a case-by-case basis, including when AI models copy others' works for the purpose of enhancing products or services that may compete with those of the content creators. In the search context, AI companies often use copyrighted works in ways that displace the original content creators, such as copying copyrighted works to ground their AI outputs. Such conduct constitutes copyright infringement, and there is no fair use in siphoning others' work for commercial and competitive advantage.

[1] Davis, Wes, *AI companies have all kinds of arguments against paying for copyrighted content* (The Verge, Nov. 4, 2023) https://www.theverge.com/2023/11/4/23946353/generative-ai-copyright-training-data-openai-microsoft-google-meta-stabilityai.

Yelp Inc., 350 Mission Street, 10th Floor, San Francisco, California 94105

About Yelp

Founded in 2004, Yelp owns and operates popular local search websites (e.g, Yelp.com) and related mobile applications, which enable users to share information about their communities. Yelp, among other things, provides and publishes a forum for members of the public to read and write reviews about local businesses, services, and other entities including nonprofits and government agencies. One of Yelp's founding principles is that the best source for information about a local community is the community members themselves. Yelp helps the public make more informed choices about local businesses and activities. As of December 31, 2022, Yelp users have contributed a total of 265 million cumulative reviews to Yelp's platform.[2]

Consumers who contribute their own content to the platform–such as ratings, reviews, photos, videos, and more–retain ownership of their individually contributed content, and grant to Yelp a license to use the content under certain circumstances. Yelp's users do not grant a license to anyone other than Yelp, and it is important to Yelp to guard against the misappropriation of its users' content and infringement of their copyrights. Yelp combines this consumer content with information contributed by businesses (e.g., hours of operation, telephone numbers, or descriptions and photos of their business, products and services). Yelp also contributes its own content to its platform, including visual interfaces, interactive features, aggregate ratings for individual businesses, and other elements and components of the service. Yelp believes that popular generative AI models, including those operated by dominant companies like Google, are using content owned by Yelp and consumers without their permission.[3]

---

[2] Yelp Internal Data, 2022. Contributed reviews include those that are recommended, not recommended, or removed from Yelp's platform.

[3] Schaul, Kevin et al., *Inside the secret list of websites that make AI like ChatGPT sound smart* (Washington Post, April 19, 2023) https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

Responses to the Copyright Office's Specific Questions

8. Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use? Please discuss any case law you believe relevant to this question.

The unauthorized use of copyrighted works to train AI models should only constitute fair use under limited circumstances, when the model's use is either clearly non-expressive or otherwise transformative. Regulators and the courts also should appreciate that AI models may use copyrighted works for different purposes–the use of the same copyrighted work may be non-expressive in one application, but expressive and not transformative in another.

Courts have generally agreed that copying without permission is fair use in the context of non-expressive intermediate copying, when the purpose is to discover unprotectable information or as a minor step towards developing an entirely new product. *See, e.g., Sony Computer Ent. v. Connectix Corp.*, 203 F.3d 596, 608 (9th Cir. 2000) (fair use to copy Sony's software to reverse engineer it and create a new gaming platform); *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 644–45 (4th Cir. 2009) (fair use to copy student essays for use in plagiarism detection software). Developers of generative AI models have used this body of law to defend their unauthorized use of copyrighted works to train AI models, including by arguing that the information loss purportedly inherent in reducing the training data down to a model through machine learning eliminates the expressive use of copyrighted works.

But expressive uses occur and persist. For example, some AI systems are capable of essentially memorizing copyrighted works in training data and communicating the original expression in response to prompts.[4] In addition,

[4] *See also* Gowthami Somepalli et al., Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, in ARXIV (2022), https://arxiv.org/abs/2212.03860; Nikhil Kandpal et al., Deduplicating Training Data Mitigates Privacy Risks in Language Models, in ARXIV (2022), https://arxiv.org/abs/2202.06539

even when the response does not communicate the original expression, AI systems are also capable of using competitors' data sets to improve their own expressive AI responses, which is something Yelp has long observed in the internet search context. The resulting AI responses act as market substitutes for the original content, since the user has no need to visit the competitor's platform. Fair use should not apply in these circumstances and the developers of the AI model and system should be liable for copyright infringement, particularly when the purpose of training the model is to commercialize it as a substitute for the copyright holder's work. As courts grapple with these questions, it will be necessary for the factfinder to interrogate the "precise nature" of the accused infringer's actions. *See Thomson Reuters Enter. Ctr. Gmbh v. Ross Intel. Inc., N*o. 1:20-cv-613-SB, 2023 U.S. Dist. LEXIS 170155, at *24-25 (D. Del. Sep. 25, 2023) (leaving for the jury to decide whether Ross used its AI to replicate and reproduce the creative drafting done by Westlaw's attorney-editors).

9. Should copyright owners have to affirmatively consent (opt in) to the use of their works for training materials, or should they be provided with the means to object (opt out)?

An opt in regime, rather than a system of opting out, should be instituted when it comes to copyrighted works being used as inputs to train generative AI models.

Unlike internet search engines–which typically allow web publishers to opt out of crawls if the publishers choose not to be indexed by the search engine–generative AI models are capable of hallucination and misrepresentation of the copyrighted works that are used to train them. For example, a recent peer-reviewed article in Cureus, a medical science journal, reported that Google Bard generated fictional, erroneous, or unsubstantiated information in response to certain medical queries.[5] Independent research by

[5] Kumar M, Mani U, Tripathi P, et al. (August 10, 2023) *Artificial Hallucinations by Google Bard: Think Before You Leap*, Cureus 15(8): e43313. doi:10.7759/cureus.43313,

Vectara, a company founded by former Google employees, suggests that industry-leading chatbots invent information between 3% of the time (in the case of OpenAI) and 27% of the time (in the case of Google).[6]

Hallucinations are a problem for the owners or licensees of copyrighted works, particularly when the chatbots purport to attribute the invented information to some other source. Yelp, for example, takes an aggressive stance against misleading reviews and is proud of its content moderation innovations, which include issuing Consumer Alerts related to deceptive review activities–warning messages that Yelp places over the review section of Yelp business pages when Yelp determines that there have been egregious attempts to deceive consumers, with a link to view the evidence when available. Yelp views its aggressive stance and innovations as competitive advantages. Other review sites, like Google, do not take such steps to protect consumers.

One type of Consumer Alert is the Questionable Legal Threats Alert, which Yelp uses when it receives evidence that a business may be making a dubious legal threat against a reviewer or using a contractual gag clause to prevent critical reviews. For example, Yelp posted this type of Alert to the Yelp business page of a notorious plastic surgeon to warn consumers about his questionable legal tactics that are intended to suppress criticism.[7] But AI chatbots that are asked about reviews for the plastic surgeon (and/or Yelp reviews for the plastic surgeon) do not typically reference the Alert and may only describe positive reviews, misleading consumers and potentially damaging Yelp's reputation.

---

https://www.cureus.com/articles/176775-artificial-hallucinations-by-google-bard-think-before-you-leap#!/.

[6] Metz, Cade, *Chatbots May 'Hallucinate' More Often Than Many Realize* (New York Times, Nov. 16, 2023) https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html.

[7] The plastic surgeon has also had his license suspended and later pleaded guilty to health care fraud.

An opt out regime unfairly foists an obligation on the authors or licensees of copyrighted works to police AI models for potential hallucinations or misrepresentations of the copyrighted works, and to opt out accordingly. Further, opt outs may be created that only remove copyrighted work from some, but not all, AI features. For example, Google has claimed that "Google-Extended" is a control that gives web publishers the power to manage whether their sites are used by Bard or Google's Vertex AI generative APIs, but Google's opt out does not apply to all of Google's AI applications—indeed, a total opt out from Google's use of content for its AI tools would require opting out of Google search altogether.[8]

Another drawback of an opt out regime is that it may also enable AI companies to retain, use and benefit from copyrighted content that they appropriate before a copyright holder opts out—content that they would not obtain in the first place under an opt in regime. That obligation is particularly unfair to individual content creators who have neither the time or money to take on such an obligation. Yelp believes that the obligation to ascertain whether the author of a copyrighted work consents to that work being used in training an AI model should lie with the developer of the model, and there is no compelling reason why collection of data for AI training purposes should be subject to the same industry standards as crawling websites for the purpose of search engine indexing.

15. In order to allow copyright owners to determine whether their works have been used, should developers of AI models be required to collect, retain, and disclose records regarding the materials used to train their models? Should creators of training datasets have a similar obligation?

[8] Coster, Helen, *As Google pushes deeper into AI, publishers see fresh challenges* (Reuters, Oct. 19, 2023) https://www.reuters.com/technology/google-pushes-deeper-into-ai-publishers-see-fresh-challenges-2023-10-19/.

Any developer of an AI model that intends to argue that its use of copyrighted works as training data constitutes a "fair use" should be required to keep records that are sufficient to identify the copyrighted works and their origin. The obligation should apply to all works copied as part of the potential set of training data, regardless of whether the developer believes the works were actually used to train the AI model or not.

Similarly, these developers should also be required to publicly disclose summary information about the composition of copyrighted works in the training data associated with each new publicly released AI model. Yelp recognizes that developers may have trade secret or privacy concerns about disclosing the entire contents of its training data to the public, and Yelp does not object to such concerns if they are legitimate. The summary information, however, should be enough to allow others to easily determine whether their works were part of the training data for a given model.

22. Can AI-generated outputs implicate the exclusive rights of preexisting copyrighted works, such as the right of reproduction or the derivative work right? If so, in what circumstances?

Yes, AI-generated outputs can implicate the exclusive rights of pre-existing copyrighted works when the copyright owner can prove a valid copyright and copying of the constituent elements of the work that are original.

The developers of AI models sometimes defend their use of copyrighted works for training data, and their resulting outputs, as inherently in the public interest. Google, for example, has claimed that "American law supports using public information to create new beneficial uses" and that using copyrighted works (so long as the works are publicly available, apparently) to train Bard is a new beneficial use.[9] But this defense often ignores the powerful counter

[9] Agius, Nicola, *Google sued for allegedly stealing content, data to train AI products* (Search Engine Land July 12, 2023) https://searchengineland.com/google-sued-content-data-ai-429334.

incentive that unfettered copying of copyrighted works, either for use in training or as AI-generated outputs, would create: it would reduce the incentives for copyright owners to create the content that is critical to machine learning.

The United States Supreme Court has long recognized that "the Framers intended copyright itself to be the engine of free expression. By establishing a marketable right to the use of one's expression, copyright supplies the economic incentive to create and disseminate ideas." Where "there is a fully functioning market" to encourage the creation and dissemination of a work, "permitting 'fair use' to displace normal copyright channels disrupts the copyright market without a commensurate public benefit." *Harper & Row, Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539, 568 n. 9 (1985), *see also Princeton Univ. Press v. Mich. Document Servs., Inc.*, 99 F.3d 1381, 1391 (6th Cir. 1996) (no fair use where copying would "have a deleterious effect upon the incentive" to publish the copyrighted work).

The reduction in the incentive to create that the Supreme Court has observed is likely to be more pronounced in instances where the copyright owner must *also* guard against the risks that an AI model may distort or misrepresent what the copyright owner has authored, particularly if the distortion or misrepresentation is likely to damage the owner's reputation or credibility. Faced with those risks, it would be reasonable for a copyright owner to simply not create, or at least not create on the Internet, making the Internet a less useful place for all.

24. How can copyright owners prove the element of copying (such as by demonstrating access to a copyrighted work) if the developer of the AI model does not maintain or make available records of what training material it used? Are existing civil discovery rules sufficient to address this situation?

It may be difficult for copyright owners to prove the element of copying unless the developer maintains or makes available records of what training

materials it used; that is one reason why Yelp would be in favor of clear records retention and disclosure rules for developers of AI models. Existing civil discovery rules would only be sufficient once a lawsuit is filed or when the developer has a reasonable expectation that litigation may ensue. The existing rules would not be sufficient to address instances where a copyright owner may not be immediately on notice of possible infringement by the AI model.

Reply to Initial Comments

The Copyright Office received comments from some of the largest and most well-known AI companies in the world, including Google, which already holds a monopoly in general online search. Many of those comments amount to attempts to defend the unauthorized copying of others' works under any and all circumstances, as long as the works are used to train AI models. But the Copyright Office should be skeptical about those attempts, including for the following reasons.

- *Publishers may not have a meaningful attempt to opt out of being used in AI models.* In its comments, Google touts its announcement of Google-Extended, which is effectively an opt out regime."[10] But for several reasons, Google-Extended does not cure the copyright problems that are endemic to Google's development of its AI technologies. First, as described above, it should not be publishers' responsibility to opt out of the potential misuse of their content by generative AI systems, which differ in important ways from standard internet search engines–an opt in regime would be far better, particularly for commercial generative AI systems like Google's. Second, Google's opt out does not apply to all of Google's AI applications—indeed, a total opt out from Google's use of content for its

[10] Comment from Google, October 30, 2023, COLC-2023-0006-9003 at p. 8, https://www.regulations.gov/comment/COLC-2023-0006-9003.

AI tools would require opting out of Google search altogether.[11] Third, Google apparently intends to continue to hold any third-party publisher content it currently uses for AI training purposes, and it has already benefited from using that content to improve its own competitive AI features without publishers' authorizations.

- *Chatbots that copy web publishers' content and serve it to consumers can compete with the web publishers as a market substitute because they siphon traffic that would otherwise go to the web publishers.* In its comments, Anthropic argues that the use of copyrighted works by Claude, its AI conversational interface, constitutes fair use in part because Clause "is 'a wholly new product' relative to the original work."[12] But that is not true in all instances for all chatbots. Yelp has encountered several instances where content from its platform was copied in chatbots' output–sometimes with attribution, sometimes without attribution, and sometimes with misattribution. In those instances, no consumer would have any incentive to visit Yelp for the unique compilation of business reviews, photos, tips and other content that Yelp provides, even if that information provides vital context for the content that was misappropriated, because the chatbot has already provided Yelp's content for them and doesn't necessarily inform them that it would be beneficial to visit Yelp's platform for more, thereby replacing Yelp and harming the market for or value of the original content. *See, e.g., Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258, 1290 (2023) (Gorsuch, J., concurring)(the fourth fair-use factor inquiry "requires courts to ask whether consumers treat a challenged use 'as a market replacement' for a copyrighted

[11] Swartz, Barry, *Google-Extended does not stop Google Search Generative Experience from using your site's content* (Search Engine Land Oct. 9, 2023) https://searchengineland.com/google-extended-does-not-stop-google-search-generative-experience-from-using-your-sites-content-433058.

[12] Comment from Anthropic PBC, October 30, 2023, COLC-2023-0006-9021, at p. 8, https://www.regulations.gov/comment/COLC-2023-0006-9021.

work or market complement that does not impair demand for the original.")

- *The argument that, in instances of alleged indirect copyright infringement, AI models have substantial non-infringing uses is overstated and should be proven in court.* In its comments, Meta concedes that it is possible to extract large portions of underlying copyrighted works or "otherwise harmful" outputs from generative AI models, but puts the blame in those instances on user queries, not on the AI developers.[13] Whether or not a specific AI model is capable of substantial non-infringing uses is irrelevant, however, in instances where the AI developer is engaged in direct copyright infringement. *See Spanski Enters. v. Telewizja Polska, S.*A., 434 U.S. App. D.C. 326, 335, 883 F.3d 904, 913 (2018)(substantial non-infringing use "beside the point" when distributor of a video-on-demand system used the system to communicate infringing performances). And in cases of alleged contributory infringement, AI companies should be put to the rigors of proving substantial non-infringing uses, just as any other accused infringer would be. In Yelp's anecdotal experiences, at least some AI models fall far short of actually and regularly producing new, distinctive works.

Conclusion

Yelp supports responsible and reasonable development of artificial intelligence technologies. Indeed, Yelp has developed its own AI-powered tools and features to enhance the search functionality of its platform. But there is no reason why the collection and use of others' copyrighted work for use in AI models should be immunized from copyright law merely because generative AI is a new, emerging technology. Instead, copyright law should protect creators and copyright holders from substitutional AI outputs that compete directly with their content without their permission.

[13] Comment from Meta Platforms Inc., October 29, 2023, COLC-2023-0006-9027 at p. 18, https://www.regulations.gov/comment/COLC-2023-0006-9027.