Written Response to Questions from the U.S. Senate Committee on the Judiciary Subcommittee on Crime and Counterterrorism

Hearing on Too Big to Prosecute?: Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI Training

Aug. 5, 2025

Edward Lee
Professor of Law
Santa Clara University School of Law

Dear Chair Hawley, Ranking Member Durbin, and Committee Members:

I received the following questions from Senator Klobuchar (in italics) and offer my responses below.

AI generated news summaries created from real-time scraping of journalistic sources—sometimes circumventing paywalls and violating terms of service—are not highly transformative and significantly devalue the market for the reporting necessary to make the news. These circumstances may indicate that real-time scraping of this nature may not fall under the fair use exception to copyright infringement.

This important question is raised in ongoing litigation, including the New York Times' and other news media's copyright lawsuits against OpenAI and Microsoft in the Multi-District Litigation and against other AI companies in other lawsuits.[1] It relates to the deployment of retrieval augmented generation (RAG) with AI generators to enable them to incorporate real-time information from the Internet (that was not contained in the datasets used to train the AI models).[2]

[1] *Do you believe that AI-generated news summaries based on copyrighted news reports infringes on copyrighted news content?*

As with most questions of infringement and fair use, I believe the answer will be fact specific. Consequently, I believe it is premature and unwarranted to conclude that AI-generated news summaries are "not highly transformative" or that they "significantly devalue the market for the reporting" without consideration of the evidence both sides will present in the ongoing litigation. For example, different AI generators or chatbots may differ in not only their outputs, but how those outputs are presented to the public, including how links to sources are displayed.[3] A new technology that significantly improves people's ability to conduct research, find relevant facts,

[1] *See, e.g.,* Second Am. Compl. ¶¶ 108-23, 186, New York Times Co. v. Microsoft Corp., No. 1:23-c-v-11195-SHS (May 28, 2025); Second Am. Compl. ¶¶ 73-79, 105-09, Dow Jones & Co. v. Perplexity AI, Inc., No. 1:24-cv-078984-KPF (Jan. 28, 2025). The full list of all copyright lawsuits is provided at Master List of Lawsuits v. AI, CHATGPT IS EATING THE WORLD (updated (Jun. 30, 2025).
[2] Kim Martineau, *What is retrieval-augmented generation?*, IBM (Aug. 22, 2023).
[3] For Google's AI mode, see Eugene Levin, *How Google's AI Mode Compares to Traditional Search and Other LLMs [AI Mode Study]*, SEMRUSH (Jun. 24, 2025).

and gain greater access to information serves the Copyright Clause's overriding goal of advancing knowledge in the United States.[4] And an AI generator that is multi-purpose and multi-functional—designed not simply for news summaries—will likely implicate other important considerations for promoting progress in the United States, such as making more accessible new creative tools to a larger segment of the population, including people with disabilities.[5]

With that caveat in mind, I believe the starting point is that facts themselves are not copyrightable. Original expression is. As the Supreme Court explained in the seminal case *Feist*: "The most fundamental axiom of copyright law is that '[n]o author may copyright his ideas or the facts he narrates.'"[6] Facts, including the "news of the day," "are part of the public domain available to every person."[7] This fact-expression dichotomy, along with the idea-expression dichotomy, serves important First Amendment values in our democracy, enabling widespread dissemination of facts and ideas.[8] As Judge Miner explained, "the freedom of access to facts and ideas is the history of democracy."[9] Indeed, as the Supreme Court admonished, the "[First] Amendment rests on the assumption that the *widest possible dissemination of information* from diverse and antagonistic sources is essential to the welfare of the public…."[10] Copyright promotes this First Amendment goal by leaving all facts in the public domain.[11]

Applying the fact-expression dichotomy, copyright law's most fundamental axiom, we must distinguish between (1) copying merely facts, which is not infringement, and (2) copying original expression, which is infringing if substantially similar and not a fair use. Thus, a news summary may be an infringing abridgment of a prior news article if the summary copied and included the *original expression* from the prior article in the summary.[12] However, if the news

---

[4] The AI company's development *of the model* serves a highly t*ransformative purpose, including in* enhancing people's ability to find relevant information and facts potentially more effectively. *See generally* Tim Keary*, Survey: 83% of users prefer AI search over 'traditional' Googling*, INNOVATING WITH AI (Jul. 1, 2025) (poll of IWAI's audience found more than 83% found AI search more efficient for getting answers to questions than traditional search); Golan v. Holder, 565 U.S. 302, 324 (2012) (interpreting the Copyright Clause's reference to the "progress of science" as "refer[ring] broadly to 'the creation and spread of knowledge and learning.'").

[5] *See* Edward Lee, *Fair Use and the Origin of AI Training*, 63 HOU. L. REV. (forthcoming 2025) (manuscript at pp. 190-191) (AI tools offer greater accessibility to creative pursuits for people with disabilities). There is an important distinction between an AI company's development of an AI model and a user's use of an AI generator deploying the model. An AI company's use of copyrighted works to develop an AI model may be highly transformative. See Bartz v. Anthropic PBC, -- F. Supp. 3d --, 2025 WL 1741691, at *7 (N.D. Cal. Jun. 23, 2025); Kadrey v. Meta Platforms, Inc., -- F. Supp. 3d --, 2025 WL 1752484, at *9 (N.D. Cal. June 25, 2025). By contrast, a user's use of an AI generator model might produce a short summary of news—let's say, Congress's recent passage of the GENIUS Act—without copying any copyrighted expression from other sources. Even though the AI-generated article might be simple and merely recount facts (without original expression) from other sources, the article is non-infringing and needs no fair use defense to escape liability. And that simple article a user created using AI would not vitiate the highly transformative purpose of the AI company in developing and training the new model. AI models escaped researchers' successful development for decades. *See* Lee*,* supra*,* 63 HOU. L. REV. *(manuscript at pp. 152-55*)

[6] Feist Publns., Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 344-45 (quoting Harper & Rose, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 556 (1985)).

[7] *Id.* at 348 (internal citation omitted).

[8] *See* Mary Sarah Bilder, *The Shrinking Back: The Law of Biography*, 43 STAN. L. REV. 299, 313-17 (1991).

[9] Roger J. Miner, *Exploiting Stolen Text: Fair Use or Foul Play?*, 37 J. COPYRIGHT SOC'Y 1, 10 (1989).

[10] Associated Press v. U.S., 326 U.S. 1, 20 (1945) (emphasis added).

[11] *See* Eldred v. Ashcroft, 537 U.S. 186, 219 (2003) ("As we said in *Harper & Row*, this 'idea/expression dichotomy strike[s] a definitional balance between the First Amendment and the Copyright Act *by permitting free communication of facts* while still protecting an author's expression.'") (quoting Harper & Row v. Nation Enters., 471 U.S. 539, 556 (2003) (emphasis added).

[12] *Cf., e.g.*, Barclays Capital Inc. v. Theflyonthewall.com, 700 F. Supp. 2d 310, 328 (S.D.N.Y. 2010) (defendant "no longer disputes, however, that it infringed the copyrights in these seventeen reports" summarizing plaintiffs' financial news content), *rev'd in part on other grounds*, 650 F.3d 876, 880 (2d Cir. 2011) ("Although *the extent to which the Firms'* success on the

summary did not copy any copyrightable expression, but *simply copied facts*, the news summary would not be infringing. Facts are in the public domain—and are free for all to use.[13] Moreover, AI models that are trained to identify merely the unprotected facts from sources without republishing their protected expression has a transformative purpose in using copies of articles to be able to identify the unprotected elements of the works so people can find relevant information, thereby serving the First Amendment interest in the *widest possible dissemination of information*.[14]

That a news summary is generated by AI does not change this analysis. Imagine that a second newspaper wrote a news summary based on an article first reported by the *Washington Post*. If the second newspaper merely copied facts reported by the *Post*, there is no copyright infringement. (Journalistic norms would typically require attribution to the source.) The answer would be the same if the second newspaper article were instead an AI-generated summary.

The Second Circuit's holding in *Hoehling v. Universal City Studios, Inc.* demonstrates this fundamental principle of copyright law. The court held that Michael MacDonald Mooney's and Universal City Studio's unauthorized copying of the "sabotage" interpretation of the Hindenburg's demise offered by A.A. Hoehling's book did not infringe Hoehling's copyright.[15] "Such an historical interpretation, whether or not it originated with Mr. Hoehling, is not protected by his copyright and can be freely used by subsequent authors," the court concluded.[16] "The rationale for this doctrine is that the cause of knowledge is best served when history is the common property of all, and each generation remains free to draw upon the discoveries and insights of the past."[17] It is of no moment that *Hoehling* involved historical facts while news articles typically involve recent facts at the time of publication. As the Supreme Court admonished in *Feist*, "The same is true of all facts—scientific, historical, biographical, and news of the day. '[T]hey may not be copyrighted, and are part of the public domain available to every person.'"[18]

Indeed, this fundamental principle was recognized in the "hot news" case, *INS v. AP*, in which the Court stated:

> [T]he news element—the information respecting current events contained in the literary production—is not the creation of the writer, but is a report of matters that ordinarily are *publici juris*; it is the history of the day. It is not to be supposed that the framers of the Constitution, when they empowered Congress "to promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries." (Const. Art. I, § 8, par. 8), intended to confer

---

copyright claims has alleviated their overall concerns is not clear, their victory on these claims is secure: Fly has not challenged the resulting injunction on appeal.")

[13] *See* Zalewski v. Cicero Builder Dev., Inc., 754 F.3d 95, 102 (2d Cir. 2014) ("Everything else in the work, the history it describes, the facts it mentions, and the ideas it embraces, are in the public domain free for others to draw upon. It is the peculiar expressions of that history, those facts, and those ideas that belong exclusively to their author.").

[14] *Cf.* Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596, 600 (9th Cir. 2000); Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1527-28 (9th Cir. 1992).

[15] Hoehling v. Universal City Studios, Inc., 618 F.2d 972, 974-77, 978-79 (2d Cir. 1980).

[16] *Id.* at 979.

[17] *Id.* at 974.

[18] Feist Publns., Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 348 (internal citation omitted).

> upon one who might happen to be the first to report a historic event the exclusive right for any period to spread the knowledge of it.[19]

My conclusion is further supported by the long line of cases recognizing fair use to create search engines (summarized in my written testimony in Appendices B and C), technologies that help people find relevant information online.[20] When an AI generator helps people in the United States find facts and information, that furthers the goal of advancing knowledge under the Copyright Clause.[21] As long as the summary or output of an AI generator does not copy original expression from online news sources, but copies merely facts, the dissemination of such facts does not produce a cognizable harm under Factor 4 of fair use in "*the protected aspect*" of the underlying work, to borrow Judge Leval's apt analysis in *Authors Guild v. Google*, another important technology fair use case.[22]

But, as I explained in my written testimony, the training of an AI model that routinely produces infringing outputs—such as infringing abridgments or summaries of works—due to inadequate guardrails will not likely be a fair use.[23] Under Factor 3 of fair use, it uses more of the works than is reasonably necessary for the transformative purpose it was intended.

There is also a very narrow claim of state law misappropriation of "hot news" that is not preempted by the Copyright Act. But, under the Second Circuit's five-factor test, it does not apply if the defendant did not publish the hot news "as its own" reporting but, instead gave attribution to the original source.[24] Thus, if an AI-generated article provided links to the sources of any hot news, it would not constitute misappropriation.

Whatever the outcome of the ongoing copyright litigation brought by news media against AI companies, it is also important to bear in mind copyright law does not preclude voluntary measures undertaken by relevant parties. For example, while prevailing in its fair use defense with respect to Google caching search, image search, and Google Books,[25] Google also established a partnership program, with paid licensing to news publishers, to feature their news in a Google News Showcase.[26] Contrary to common fallacy, fair use and licensing are not mutually exclusive.[27]

---

[19] INS v. AP, 248 U.S. 215, 234 (1918).
[20] *See* Edward Lee, Testimony before the U.S. Senate Committee on the Judiciary Subcommittee on Crime and Counterterrorism 14-15 (Jul. 16, 2025) (Appendices B and C).
[21] U.S. CONST. art. I, § 8, cl. 8.
[22] Authors Guild v. Google, Inc., 804 F.3d 202, 224 (2d Cir. 2015) (emphasis in original).
[23] *See* Lee, *supra* note 20, at 2, 4-5.
[24] *See* Barclays Capital Inc. v. Theflyonthewall.com, 650 F.3d 876, 903 (2d Cir. 2011) (explaining NBA v. Motorola, Inc. 105 F.3d 841, 898 (2d Cir. 1997) and INS v. AP, 248 U.S. 215, 239 (1918)).
[25] *See* Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) (Google Book search was fair use); Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007) (Google image search was fair use); Field v. Google, Inc., 412 F. Supp. 2d 1106 (D. Nev. 2006) (Google search of cached copy of website was fair use).
[26] *See, e.g.*, Sundar Pichai, Our $1 billion investment in partnerships with news publishers, GOOGLE (Oct. 1, 2020).
[27] *See, e.g.,* Sega v. Accolade, SEGA RETRO, https://segaretro.org/Sega_v._Accolade ("The two companies reached an out of court settlement which allowed Accolade to continue building their own Mega Drive cartridges, *but as an official licensee.");* Campbell v. Acuff-Rose Music, Inc., WIKIPEDIA *("*On remand, the parties settled the case out of court. According to press reports, under terms of the settlement, Acuff-Rose dismissed its lawsuit, and 2 Live Crew agreed to license the sale of its parody of the song*.").*

[2] *Do you believe that circumventing paywalls and ignoring terms of service to secure content for AI models shares similarities with downloading pirated books?*

Before discussing fair use, it is important to recognize that circumventing paywalls and ignoring terms of service are both addressed by other laws more directly tailored to such conduct. For example, circumventing a paywall to copyrighted content may violate the DMCA anti-circumvention provision.[28] (Relatedly, the Librarian of Congress has recognized a limited exception under its Section 1201 rulemaking authority for text data mining for scholarly research and teaching.[29]) Circumventing a paywall to a website may also violate the Computer Fraud and Abuse Act.[30] Finally, violating terms of service can raise a breach of contract claim.[31] Thus, regardless of how courts weigh the fair use analysis, other laws might more directly address the issue of scraping of online content in contravention of a paywall or terms of use—and create liability for such conduct.

As to how courts should weigh such conduct under fair use, my analysis is the same as my recommendation elaborated in my testimony with respect to use of pirated books.[32] I agree with Judge Chhabria's flexible approach in *Kadrey v. Meta*, in which he ruled that Meta's downloading of copies from shadow libraries was for the highly transformative purpose of training its AI model, but that such use could weigh against fair use if the plaintiffs establish market harm from such downloading.[33] But an unlawfully acquired or possessed copy should not be treated as a per se disqualification of a defendant's ability to raise a defense of fair use.

This fact-specific approach to the fair use analysis is consistent with the text of the fair use provision in the Copyright Act, which does not include any per se requirement for a "lawfully made copy" or a "lawfully possessed copy" as it does for other copyright exceptions.[34] Under a well-established canon of construction, the fair use provision should not be read to impose a limitation Congress expressly included in other copyright exceptions (such as in Section 109), but left out of the fair use provision (Section 107).[35] As Chief Justice Roberts explained for a unanimous Court in an analogous situation involving a notice exception of the Tax Code that

---

[28] *See* 17 U.S.C. § 1201(a)(1); *see also* Theresa M. Troupson, Note, *Yes, It's Illegal to Cheat a Paywall: Access Rights and the DMCA's Anticircumvention Provision*, 90 N.Y.U. L. REV. 325, 350-52 (2015).

[29] *See* 37 CFR 201.20; *Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies*, FED. REG. (Oct. 28, 2024).

[30] *See* hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1198 (9th Cir. 2022) ("*Van Buren*[, 593 U.S. 374 (2021)] stated that the CFAA's password-trafficking provision, section 1030(a)(6), which also uses the word 'authorization,' 'contemplates a 'specific type of authorization—that is, authentication,' which turns on whether a user's credentials allow him to proceed past a computer's access gate, rather than on other, scope-based restrictions."); 18 U.S.C. § 1030(a)(2)(C) ("[w]hoever ... intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains ... information from any protected computer ... shall be punished" by fine or imprisonment.").

[31] Breach of contract is the first claim in Reddi*t's suit against Anthr*opic for alleged unauthorized scraping in violation of the terms of service. *See* Complaint, Reddit, Inc. v. Anthropic, PBC, No. CGC-25-62582 (Jun. 4, 2025).

[32] *See* Lee, *supra* note 20, at 5-8.

[33] *See* Kadrey v. Meta Platforms, Inc., -- F. Supp. 3d --, 2025 WL 1752484, at *12, *21 (N.D. Cal. June 25, 2025). I disagree with Judge Chhabria's endorsement of a new theory of market dilution in dicta, however. *See* Lee, *supra* note 20, at 9-12.

[34] Compare 17 U.S.C. § 109(a) ("the owner of a particular copy … lawfully made under this title") (emphasis added) with id. § 107 ("fair use of a copyrighted work"); Kirtsaeng v. John Wiley & Sons, Inc., 568 U.S. 519, 537 (2013) (discussing "lawfully made" copy requirement in §§ 109(c) (exception to public display), 109(e) (exception for video games in coin-operated equipment), and 110(1) (in-classroom teaching exception to public display and performance but not if copy "not lawfully made"); see also 17 U.S.C. § 108(c)(2) ("lawful possession of such copy*" by library or archives*) (emphasis added).

[35] Sebelius v. Closer, 569 U.S. 369, 378 (2013) ("We have long held that '[w]here Congress includes particular language in one section of a statute but omits it in another section of the same Act, it is generally presumed that Congress acts intentionally and purposely in the disparate inclusion or exclusion.'") (quoting Bates v. United States, 522 U.S. 23, 29-30 (1997)).

lacked a requirement expressly contained in a following section, "Had Congress wanted to include a legal interest requirement, it certainly knew how to do so. The very next provision—also enacted as part of the Tax Reform Act of 1976—requires the IRS to" follow such a requirement.[36] This same principle applies with equal force here to the Copyright Act of 1976. Section 109(a) imposes a requirement of a "lawfully made copy," but Section 107 does not.

The fact-specific approach to a defendant's initial acquisition of a copy that was unlawfully made is also consistent with the Supreme Court's repeated admonition that fair use is fact-specific and has no bright-line rules.[37] The Supreme Court did not treat as dispositive the *purloined* nature of a manuscript in *Harper & Row*, and, in *Google v. Oracle*, the Court rejected the argument that courts should consider the "bad faith" of the defendant under fair use, preferring instead the view of Judge Leval's influential article recognizing that "[c]opyright is not a privilege reserved for the well-behaved."[38]

This is not to suggest that defendants have a green light to do whatever they want under fair use. They do not. For example, a defendant's circumventing paywalls may provide evidence of cognizable market harm under Factor 4 of fair use in some cases, especially if wide-scale. The overarching point is that courts are well-equipped to weigh all these considerations and the evidence presented by the parties on a case-by-case basis.

Finally, just as with my answer to the first question above, we must recognize that voluntary practices related to scraping of online content are already developing. Many AI companies, including OpenAI, Amazon, Google, and Microsoft,[39] have voluntarily agreed to follow the EU's General Purpose AI Code of Practice. Under this Code of Practice, companies agree to use scraping of "only lawfully accessible copyright-protected content" and "not to circumvent effective technological measures" protecting copyrighted content.[40]

Given the more than 40 copyright lawsuits pending before the courts, other laws that directly address issues related to paywall circumvention and the breach of terms of service, and the development of voluntary practices among AI companies related to scraping of online content, I believe there is no need for Congress to intervene.

---

[36] Polselli v. IRS, 598 U.S. 432, 439 (2023).
[37] Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 528 (2023); Google LLC v. Oracle Am., Inc., 593 U.S. 1, 18-19 (2021); Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 577 (1994).
[38] Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 563 (1985); Google LLC v. Oracle Am., Inc., 593 U.S. 1, 32-33 (2021) (quoting Pierre N. Leval, Toward a Fair Use Standard, 103 HARV. L. REV. 1105, 1126 (1990)).
[39] *See Signatories Code of Practice*, EC EUROPA. For more on the development of the Code of Practice, see *Drawing-up a General-Purpose AI Code of Practice*, EUR. COMM'N.
[40] *Code of Practice for General-Purpose AI Models*, Copyright Chapter measure 1.2, EUROPA, at https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai#ecl-inpage-Signatories-of-the-AI-Pact.