On Copyright, "Facts," & Generative AI

By Benjamin L. W. Sobel (Cornell Tech)

On Copyright, "Facts," & Generative AI

By Benjamin L. W. Sobel (Cornell Tech)

Essential to copyright law is the distinction between "facts" and "expression." Facts are uncopyrightable; as the Supreme Court has explained, facts are *discovered* rather than *authored*.[1] Authorial expression, on the other hand, is copyrightable. This legal rule ensures, at least in theory, that the public can freely use facts about the world, while authors can still secure protection for their original creations.[2] This essay considers the future of copyright's fact-expression distinction in an era of generative AI; more than anything, it memorializes in writing a comment I've found myself repeating at copyright and tech-law conferences all year.

The distinction between unprotectable facts and protectable expression has helped useful information technologies flourish. Google Books, for example, was held to be a "fair use" that did not infringe copyright, in part because "the purpose of Google's copying of the original copyrighted books is to make available significant information *about those books*," such as where or how frequently a given word appears in a text.[3] Although Google Books made "snippets" of the books' text available for viewing, a court concluded that this display function reproduced text in such a "cumbersome, disjointed, and incomplete" manner that "it would be a rare case in which the searcher's interest *in the protected aspect* of the author's work would be satisfied by what is available from snippet view, and rarer still . . . that snippet view could provide a significant substitute for the purchase of the author's book."[4] In other words, Google Books provided *facts* about books, but in light of the tool's purpose, it did not reproduce enough *expression* from those books to infringe copyright. Using similar reasoning, courts have also held technologies like image search

[1] *See* Feist Publications, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 347 (1991).
[2] *Id.* at 347-48.
[3] Authors Guild v. Google, Inc., 804 F.3d 202, 217 (2d Cir. 2015).
[4] *Id.* at 224-25.

and plagiarism-detection software to be non-infringing even though they, like Google Books, entail large-scale reproductions of copyrighted materials.[5]

The legal scholar Matthew Sag has identified in these cases a doctrinal principle that he calls "non-expressive use."[6] Sag explains that "in general, the copyright owner's exclusive rights are limited to the right to communicate the expressive aspects of her work to the public. To put it another way, copyright typically only concerns itself with the threat of expressive substitution."[7] "[T]echnical acts of copying that do not communicate the original expression" of the copied work "to a new audience do not interfere with the interest in original expression that copyright is designed to protect" and thus do not infringe copyright.[8] Sag argues that generative AI—which may train on innumerable unauthorized copies of copyrighted works in order to produce expressive outputs that resemble those works—is generally shielded from infringement liability by the same principle of non-expressive use that shielded earlier information technologies.[9]

Reasonable minds might disagree about whether generative AI is really a non-expressive use of the copyrighted works that appear in training data.[10] Sag suggests that "perhaps rare[ly] . . . the process of creating generative AI may cross the line from fair use to infringement because these large language models sometimes 'memorize' the training data rather than simply 'learning' from it."[11] I'm inclined to think that generative AI may impermissibly appropriate the expressive value in its training data much more consistently.[12] (Part of our disagreement stems from my belief that training AI on copyrighted works without authorization may infringe copyright even when the AI does not generate output "substantially similar" to works in its training data—just as a human

---

[5] *See generally* A.V. ex rel. Vanderhye v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009); Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007). *See also* Matthew Sag, *Orphan Works as Grist for the Data Mill*, 27 BERKELEY TECH. L.J. 1503, 1542 n.190 (2012).
[6] Sag, *supra* note 5, at 1528.
[7] *Id.*
[8] Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295, 306 (2023).
[9] *See id.* at 307-10.
[10] *Compare id.* with Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 68-79 (2017).
[11] Sag, *supra* note 10, at 301.
[12] *See generally* Sobel, *supra* note 10; Benjamin L. W. Sobel, *Elements of Style: Copyright, Similarity, and Generative AI*, 38 HARV. J.L. & TECH. __ (forthcoming), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4832872.

who pirated textbooks and learned from them would be liable for copyright infringement, even if she never wrote substantially similar textbooks of her own.[13] Sag, by contrast, asserts that generative AI "models still qualify as non[-]expressive use so long as the[ir] outputs are not substantially similar to any particular original expression in the training data."[14])

But this essay isn't exactly about whether today's generative AI qualifies as "non-expressive." Rather, it's about whether the underlying premise of non-expressive use will survive in the era of generative AI. My argument is not that Sag is wrong to identify a principle of non-expressive use that animates many major copyright decisions; on that score, he's certainly correct. What's less clear is whether a non-expressive use principle fully resolves generative AI's legal status.

There are a few reasons to believe that, when push comes to shove, courts and/or policymakers may limit the principle of non-expressive use and perhaps even revise the fact-expression distinction as we know it. Consider the Google Books case. The court there was clear that the utility of Google Books did not derive from "the protected aspect of the author[s'] work[s]"—Google Books was, by and large, furnishing facts about books.[15] To the extent Google Books diminished demand for books it allowed users to preview, it did not do so in a way cognizable as copyright infringement, because it did not interfere with demand for "the protected aspect" of those works: their expression.

But in that same case, the court also observed,

> Google does not provide snippet view for types of books, such as dictionaries and cookbooks, for which viewing a small segment is likely to satisfy the searcher's need. The result of these restrictions is, so far as the record demonstrates, that a searcher cannot succeed, even after long extended effort to multiply what can be revealed, in revealing through a

[13] *See* Benjamin L.W. Sobel, *Copyright Accelerationism*, 100 CHI.-KENT L. REV. __ (forthcoming 2025), manuscript at 23, *available at* https://ssrn.com/abstract=4658701.
[14] Sag, *supra* note 10, at 309.
[15] *Authors Guild*, 804 F.3d at 224-25 (emphasis omitted).

> snippet search what could usefully serve as a competing substitute for the original.[16]

As I observed in 2017, there is some tension in the court's reasoning here.[17] If the unprotectable elements of a work truly are free for the taking, irrespective of how it might affect the market for that work, then Google should have been free to reproduce segments of cookbooks that would probably "satisfy [a] searcher's need" on their own. After all, the principal reason we read cookbooks is to access lists of ingredients and step-by-step recipes, which may generally be uncopyrightable facts rather than protected expression.[18] (As an aside, I suspect that copyright's idea-expression distinction motivates online recipe blogs' seemingly universal practice of prefacing their recipes with interminable stories about the authors' childhoods, their family histories, and/or the arcane dietary preferences of the members of their households. These walls of text are copyrightable insulation for recipe sites' main draw: uncopyrightable recipes.)

This same phenomenon appears in an earlier decision from the same court that decided the Google Books case. In *American Geophysical Union v. Texaco*, the Second Circuit court of appeals held that it was not fair use for a scientist in Texaco's research division to photocopy copyrighted journal articles in order to "facilitate . . . current or future professional research."[19] The court noted explicitly that the copied journal articles were "essentially factual in nature," and, moreover, that the evidence indicated that a representative Texaco researcher "was interested exclusively in the facts, ideas, concepts, or principles contained within the articles."[20] The court elaborated: "Though scientists surely employ creativity and originality to develop ideas and obtain facts and thereafter to convey the ideas and facts in scholarly articles, it is primarily the ideas and facts themselves that are of value to other scientists in their research."[21] Nevertheless, the *Texaco* court rejected the fair use defense, primarily because the purpose of the copying was to

---

[16] *Id.* at 222.
[17] Sobel, *supra* note 10, at 56-57.
[18] *See* What Does Copyright Protect? (FAQ), U.S. Copyright Office, https://www.copyright.gov/help/faq/faq-protect.html (last visited Jun 3, 2024) ("A mere listing of ingredients is not protected under copyright law. However, where a recipe or formula is accompanied by substantial literary expression in the form of an explanation or directions, or when there is a collection of recipes as in a cookbook, there may be a basis for copyright protection.").
[19] Am. Geophysical Union v. Texaco Inc., 60 F.3d 913, 915 (2d Cir. 1994).
[20] *Id.* at 925 & n.11.
[21] *Id.*

create more copies for which Texaco could have purchased a photocopying license from the rightsholders.

These excerpts from the Google Books and *Texaco* decisions suggest that an enterprise of extracting factual information from copyrighted works may not always be a silver-bullet defense to claims of copyright infringement. If the logic of non-expressive use threatens to undercut what we regard as a copyright owner's heartland market—like the sale of cookbooks or dictionaries to readers interested in recipes or definitions—then cases like *Authors Guild* and *Texaco* suggest that courts may hesitate to adopt that logic wholesale.

Generative AI strip-mines information from expressive works and uses that information in ways that may undermine the established markets for those works. Proponents of the non-expressive use principle for generative AI assert that the information that AI models "learn" from copyrighted works is factual and thus does not implicate the protected aspects of those works.[22] But, as the DLI's own James Grimmelmann has chronicled, information's status as "fact" or "expression" is a contested boundary.[23] At some level of generality, "the characteristic aspects of syntax, diction, and tone that make Lorrie Moore's writing so charming" are surely bare facts about how Moore strings words together; at a more particular level, "the characteristic aspects of syntax, diction, and tone that make Lorrie Moore's writing so charming" are precisely what *constitute* Lorrie Moore's authorial expression.[24] Indeed, the qualities that constitute expression in any communication are in some fundamental sense factual; ascertainable, factual qualities of written English are what make it sound "right" or "natural," and English instructors instrumentalize these facts to correct "unnatural" prose. Arguably, even reality as we know it is socially constructed.[25]

---

[22] *See*, *e.g.*, Motion to Dismiss Plaintiffs' Complaint at 2, Dkt. No. 23, Kadrey v. Meta Platforms, Inc., No. 3:23-cv-3417 (N.D. Cal. Sept. 18, 2023) ("Copyright law does not protect facts or the syntactical, structural, and linguistic information that may have been extracted from books like Plaintiffs' during training.").
[23] *See generally* James Grimmelmann, *Three Theories of Copyright in Ratings*, 14 VANDERBILT J. ENT. & TECH. L. 851 (2012).
[24] *See* Benjamin L.W. Sobel, *Elements of Style: Copyright, Similarity, and Generative AI*, 38 HARV. J.L. & TECH. __ (forthcoming), manuscript at 38-48, *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4832872.
[25] *See* Justin Hughes, *Created Facts and the Flawed Ontology of Copyright Law*, 83 NOTRE DAME L. REV. 43, 52-53 (2007).

Expression is the musculature that connects facts into intelligible communications. Those who defend generative AI against charges of copyright infringement characterize the technology as something like a master butcher or a school of piranhas: expressive works go in, flesh is exhaustively separated from bone, and distilled facts come out. But what does it mean to skeletonize, say, a biology textbook in this fashion? What sort of information product does such a process realize? A good science textbook is *already* a masterfully efficient presentation of factual material; how are we to imagine such a work being reduced to facts *alone*?[26] Maybe the facts are simply "that which remains when a work is paraphrased." (Famously, however, a federal court of appeals held that book of trivia questions about the TV series *Seinfeld*—with right and wrong answers that corresponded to "facts" about the show's fictional plot!—infringed the copyright in the TV show.[27] The defendant's effort to distill *Seinfeld* to factual multiple-choice questions still ended up reproducing copyrighted expression.[28])

But generative AI is notoriously *bad* at learning the sorts of facts that a student might learn from a history or biology textbook. This is precisely the information that an AI model cannot be trusted to reproduce accurately: models consistently "hallucinate" erroneous facts. Instead, what these models extract from training data are the qualities that make expression *seem* expressive. You can depend on today's AI chatbots to produce fluent prose, but you can't depend on them to furnish discrete facts about the world. In other words, the models aren't learning facts to express; they're learning how to synthesize expression without regard for the factual propositions expressed.

That generative AI can produce convincingly expressive output by analyzing expressive works suggests that the information it gleans from training data may resemble expression more than fact. But even if training generative AI implicates only on what copyright law would consider facts, the technology still presents a vexing question for information policy: should expression remain the law's focus? For a few hundred years, expression may have been a decent proxy for the aspects of the market for expressive works that we wanted to protect. Copying of facts was generally legitimate activity that we

[26] *See id.* at 57 (offering "a succinct rendering of the problem in copyright law: the canonical way to specify the fact is the same as the way to specify the expression of the fact, i.e., by expressing it.").
[27] Castle Rock Ent., Inc. v. Carol Pub. Grp., Inc., 150 F.3d 132, 135-36, 139 (2d Cir. 1998).
[28] *See id.*

didn't want to encumber, and copying of expression generally entailed the free-riding we thought a regulatory regime ought to forbid. But the very first American copyright statute—the Copyright Act of 1790, enacted just two years after the ratification of the Constitution—described itself as "[a]n Act for the encouragement of learning, by securing the copies of *maps, Charts, And books*, to the authors and proprietors of such copies."[29] "Maps" and "charts" are paradigmatically useful not because of the expressive choices they embody, but because of the facts they disclose. Information policy for different eras may require different emphases on facts versus expression.

A biology or history textbook primarily serves to impart factual information to its readers. Like the researcher who copied journal articles in the *Texaco* case, students read these textbooks to glean unprotectable facts and ideas. Yet there is no serious legal argument that a student could pirate a textbook simply because her purpose in doing so was to extract unprotectable facts by reading it, even if the logic of non-expressive use might suggest such a conclusion. A court would balk at this argument for the same reason a court commended Google for omitting cookbooks from Google Books: judges hesitate to undercut markets that seem like copyright's heartland. Courts and/or policymakers might similarly balk at generative AI's non-expressive use defense if the technology threatens to undermine incentives for, say, original journalism or other expressive endeavors.[30]

Authors Guild, *Texaco*, and copyright protections for maps and charts all reflect a recognition that legitimate demand for expressive works may derive to some degree from features that aren't exactly expressive. In media like scientific journal articles, recipes, and textbooks, copyrightable expression is often just a substrate for the factual information that really interests us. Historically, protecting those works' expression *qua* expression may have been a good-enough proxy for protecting what we saw as their authors' legitimate markets. But now, with a surgical precision that outpaces earlier information-processing technologies, generative AI extracts valuable, yet putatively non-expressive, information from expressive works. Generative AI technology has wrought such uncertainty for

[29] Act of May 31, 1790, 1 Stat. 124 (emphasis added), *available at* https://www.copyright.gov/history/1790act.pdf. *See also* Am. Dental Ass'n v. Delta Dental Plans Ass'n, 126 F.3d 977, 978 (7th Cir. 1997) ("Maps and globes are not only copyrightable, but also constituted two-thirds of the original scope of copyright." (citation omitted)) (Easterbrook, J.).
[30] *Cf.* William Fisher, *Integrating AI and IP: A Legal Realist Approach* at 22:00, Reframing Intellectual Property Law in the Age of Artificial Intelligence (Hong Kong University, Dec. 16, 2023), https://youtu.be/XmRM0RxjJg4?t=1319.

copyright law because it jeopardizes the industrial-policy commitments that underly copyright's fact-expression distinction.

In other words, generative AI challenges us to assess whether authorial expression—the proxy we've used to shape the market for creative works—still helps us structure a market for information as we would like it to be structured. Concluding that the authorship proxy no longer serves us certainly does not mean that we must extend today's copyright entitlements to facts writ large; doing so would disastrously inhibit the free exchange of information and contravene the Supreme Court's determination that "originality is a constitutionally mandated prerequisite for copyright protection."[31] Rather, such an insight might challenge us to craft information policy for the AI age that replaces the copyright regime as we know it—with its fact-expression distinction and its reliance on authorship as a proxy for legally protected value—with something that serves us better.

Benjamin L. W. Sobel
DLI Postdoc, Cornell Tech
bsobel@cornell.edu
More Info >

[31] *Feist*, 499 U.S. at 346, 351.