THE FREEDOM TO EXTRACT IN COPYRIGHT LAW*

MOLLY SHAFFER VAN HOUWELING**

Contemporary technological and legal developments are raising the stakes for the freedom to extract—that is, the freedom to identify and reuse the unowned ideas, facts, and methods that appear in copyrighted works. Advances in data science and artificial intelligence present remarkable opportunities to extract valuable patterns from society's vast store of existing expressive works. At the same time, some of the results of that extraction may threaten the interests of copyright owners in ways that extraction never has in the past. And those results may themselves be disseminated in ways that obscure valuable information and thus deny the public the full value of extraction. This Article introduces the concept of the freedom to extract, prescribes ways in which the law could better safeguard it, and explains how those prescriptions might apply to contemporary controversies over artificial intelligence and other novel types of extraction. I argue that we should evaluate technological extraction on the grounds of whether it extracts unprotectable material and on whether it facilitates beneficial extraction by others. Extraction need not be exploitative—in the sense of taking from the intellectual commons without contributing back—if it is accompanied by extractability.

** Harold C. Hohbach Distinguished Professor of Patent Law and Intellectual Property, University of California, Berkeley, School of Law. Thanks to BJ Ard, Oren Bracha, Rebecca Chambers, Colleen Chien, Eric Goldman, James Grimmelmann, David Hansen, Jung Kim, Mark Lemley, Jessica Litman, Peter Menell, Robert Merges, Tejas Narechania, Blake Reid, Matthew Sag, Pamela Samuelson, Erik Stallman, and participants in the Berkeley Law Faculty Workshop and the Works in Progress in Intellectual Property Conference for helpful comments. Thanks to Akshat Agrawal, Nicole Boucher, Zhudi Huang, Allen Park, Sakshi Sharma, and Sherry Shi for extraordinary research assistance. And thanks to Meyke Kang and her colleagues at the *North Carolina Law Review* for outstanding editorial work.

INTRODUCTION

Copyright protection does not extend to any "idea, procedure, process, system, method of operation, concept, principle, or discovery."[1] It covers only concrete expressions of those unprotected elements. This is the essence of the so-called "idea/expression" distinction in copyright law. This distinction reflects an implicit cost-benefit analysis: keeping unprotected elements in the public domain for others to use as building blocks ensures the iterative accretion of knowledge and creativity, while protection for specific expressions of

1. 17 U.S.C. § 102(b).

unprotected elements gives creators enough of a competitive advantage to incentivize their creativity. The soundness of this cost-benefit analysis depends on facts about the creative environment that are currently in flux due to developments in generative artificial intelligence. These developments raise the stakes for understanding—and perhaps fine-tuning—the role of the idea/expression distinction and the related doctrines that support it.

One way to understand the idea/expression distinction is as a delineation of the scope of the subject matter protected by copyright, specifying what *is* protected. The distinction can also be understood from the other side, in terms of the freedom it affords nonowners to use what copyright leaves unprotected. Specifically, it grants nonowners who encounter a copyrighted work the freedom to extract from that work the unprotected elements within it and to use them without the copyright owner's permission. This important but under-theorized *freedom to extract* is the lens through which this Article will examine the role of the idea/expression distinction in the age of artificial intelligence ("AI").

Consider a concrete example of extraction: copyright protection for a book about bicycle racing tactics prohibits other people from copying the book's text (or substantially similar variations) without permission. But copyright leaves people free to extract the tactical insights from the book, to deploy them, and to write their own books explaining them. This is the freedom to extract in action.

Although the idea/expression distinction provides a solid doctrinal foundation for this type of extraction, the doctrine's seemingly simple on/off switch can fail to meaningfully protect the freedom to extract. For example, in the course of extracting unprotected elements, subsequent authors sometimes copy expression as well. A subsequent author of a book on cycling tactics may copy the details of a diagram depicting a particular tactical ploy. The diagram may be both a concrete expression of a particular tactical idea *and* the best way to convey the essence of that idea clearly. If copyright law were to strictly forbid copying of such expression, it could make extracting unprotected elements unreasonably difficult as a practical matter.

Fortunately, copyright law is more subtle than the simple version of the idea/expression distinction might suggest. Several different doctrines within copyright law can operate to prevent liability for defendants whose copying of expression is incidental to lawful extraction of unprotected elements. The most obvious of these is the "merger" doctrine, which applies when, as in the cycling diagram example, there is such a narrow range of ways to express an idea that

courts allow copying of expression in order to safeguard the freedom to extract the idea.[2]

Although well-established and widely invoked, the merger doctrine is not alone sufficient to safeguard the freedom to extract. Over time, technological and legal developments have threatened the freedom to extract in ways that the merger doctrine does not address. These developments include copyright protection for computer software, which is typically distributed in ways that obscure unprotected elements and thereby insulate them from easy extraction. The deployment of technological tools (encryption, password protection, etc.) that can restrict access to unprotected elements poses another challenge to the freedom to extract.

Other technological developments over the past several decades have made the freedom to extract more valuable than ever and have thus raised the stakes for its protection. Specifically, new computational capacity and techniques for text data mining allow for the extraction of unprotected elements from vast bodies of expressive works. This development creates unprecedented opportunities for augmenting our knowledge about those works and, in turn, about the societies that have produced and been influenced by them. Courts have generally recognized the value of extraction in these contexts, deploying fair use and other doctrinal tools to augment the merger doctrine and thereby safeguard the freedom to extract in a changing technological and legal environment.

Contemporary technological and legal developments are again raising the stakes for the freedom to extract. Artificial intelligence presents remarkable new opportunities to extract valuable patterns from society's vast store of existing expressive works. At the same time, some of the results of that extraction may threaten the interests of copyright owners in ways that extraction never has in the past. And those results may themselves be disseminated in ways that obscure valuable information and thus deny the public the full value of extraction.

In Part I of this Article, I synthesize and build upon the work of other scholars in order to articulate and elaborate on the concept of the freedom to extract, explaining why it is important to a well-functioning copyright system.[3]

2. *See* RESTATEMENT OF COPYRIGHT § 17 (AM. L. INST., Tentative Draft No. 5, 2024) ("When protecting the particular expression of an unprotectable element in a work of authorship would have the effect of extending protection to the unprotectable element itself, then the scope of copyright does not extend to that particular expression. That expression is said to have *merged* with the unprotectable element.").

3. For other scholarship describing aspects of the freedom to extract, see, for example, Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 749–50 (2021); Pamela Samuelson, *Why Copyright Law Excludes Systems and Processes from the Scope of Its Protection*, 85 TEX. L. REV. 1921, 1933 (2007) [hereinafter Samuelson, *Why Copyright Law Excludes*]; Matthew Sag, *The New Legal Landscape*

I then explain how doctrinal protection for the freedom to extract has evolved alongside technological and legal developments—generally safeguarding the freedom to extract but occasionally falling short. In Part II, I prescribe ways in which copyright law could better safeguard the freedom to extract in general. In Part III, I explain how those prescriptions might apply to contemporary controversies over generative artificial intelligence and other novel types of extraction. I argue that understanding the freedom to extract should lead us to evaluate technological extraction on the grounds of whether it extracts unprotectable material *and* on whether it facilitates beneficial extraction by others. Extraction need not be exploitative—in the sense of taking from the intellectual commons without contributing back—if it is accompanied by *extractability*.

I. THE VALUE AND DOCTRINAL STATUS OF THE FREEDOM TO EXTRACT

This part introduces the concept and normative significance of "extractive use," by which I mean use of copyrighted works for the purpose of extracting unprotected ideas, methods, or facts. It then traces doctrinal recognition of the validity of extractive use—including many instances in which courts have recognized the importance of the freedom to extract and some cases in which they have given it short shrift.

A. *Introducing the Concept of Extractive Use*

It is a foundational principle of copyright law that protection attaches only to the expression embodied in copyrighted works, not to the underlying substance conveyed by that expression.[4] As the Supreme Court explained in *Baker v. Selden*,[5] "[t]he very object of publishing a book on science or the useful arts is to communicate to the world the useful knowledge which it contains. But this object would be frustrated if the knowledge could not be used without incurring the guilt of piracy of the book."[6] The knowledge at issue in *Baker v. Selden* was a system of bookkeeping.[7] Other early cases excluded from copyright

for Text Mining and Machine Learning, 66 J. COPYRIGHT SOC'Y U.S.A. 291, 304 (2019) [hereinafter Sag, *The New Legal Landscape*]; Paul J. Heald, *The Extraction/Duplication Dichotomy: Constitutional Line-Drawing in the Database Debate*, 62 OHIO ST. L.J. 933, 944 (2001); Jessica Litman, *Revising Copyright Law for the Information Age*, 75 OR. L. REV. 19, 33 (1996); James Grimmelmann, *Copyright, Technology, and Access to the Law: An Opinionated Primer* 27–28 (N.Y.L. Sch. Legal Stud. Rsch. Paper No. 08/09-1, 2008) [hereinafter Grimmelmann, *Opinionated Primer*], https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1156829 [https://perma.cc/XNX7-Z29U (staff-uploaded archive)]; WILLIAM F. PATRY, PATRY ON FAIR USE § 3:53 (2024).

4. "Unlike a patent, a copyright gives no exclusive right to the art disclosed; protection is given only to the expression of the idea—not the idea itself." Mazer v. Stein, 347 U.S. 201, 217 (1954).

5. 101 U.S. 99 (1879).

6. *Id.* at 103.

7. *See generally* Samuelson, *Why Copyright Law Excludes*, *supra* note 3, at 1922 (emphasizing that *Baker* was about the exclusion of *systems* from copyright protection).

protection a way of designating different zones on a map depicting fire risk in a city,[8] systems of shorthand and stenography,[9] tax and pension programs,[10] and the themes of plays.[11]

This concept—that copyright protects the particular way that knowledge is expressed but not the knowledge itself—is often referred to as the "idea/expression" distinction, although "idea" does not fully capture the elements of copyrighted works that are unprotectable.[12] *Baker v. Selden* itself can best be described as excluding protection for a *system*—not the mere abstract idea of an accounting system, but a particular way of performing accounting tasks.[13] The Copyright Act of 1976 codified the exclusions articulated in *Baker* and other cases with a long list of unprotected elements: ideas, procedures, processes, systems, methods of operation, concepts, principles, and discoveries.[14] The Supreme Court has also held that facts are not copyrightable,[15] nor are any other elements that are not original to the copyright owner.[16] These exclusions can be grouped into three rough categories: *ideas* (meaning abstract conceptions and including the statutorily excluded elements of "ideas," "concepts," and "principles"); *methods* (meaning ways of doing things and including "procedures, processes, systems, [and] methods of operation"); and *facts* (including "discoveries").[17]

These exclusions reflect a judgment that the underlying purpose of copyright—to promote the progress of knowledge ("science," to use the framers' term)[18]—would be undermined rather than served if copyright protection prohibited subsequent authors from reusing elements in any of these

8. Perris v. Hexamer, 99 U.S. 674, 676 (1878).

9. *See, e.g.*, Griggs v. Perrin, 49 F. 15, 15–16 (C.C.N.D.N.Y. 1892); Brief Eng. Sys., Inc. v. Owen, 48 F.2d 555, 556 (2d Cir. 1931).

10. Aldrich v. Remington Rand, Inc., 52 F. Supp. 732, 733 (N.D. Tex. 1942) (tax collection system); Long v. Jordan, 29 F. Supp. 287, 290 (N.D. Cal. 1939) (pension system).

11. Dymow v. Bolton, 11 F.2d 690, 692 (2d Cir. 1926); Eichel v. Marcin, 241 F. 404, 411 (S.D.N.Y. 1913); *see also* Pike v. Nicholas [1869] 5 Ch App. 251, 268 (Eng.) (denying protection to historical theories).

12. *See generally* Oren Bracha, *The Ideology of Authorship Revisited: Authors, Markets, and Liberal Values in Early American Copyright*, 118 YALE L.J. 186 (2008) (discussing the evolution of the "idea/expression" distinction). On the importance of recognizing the breadth of the exclusion of unprotectable elements beyond abstract ideas, see generally Samuelson, *Why Copyright Law Excludes*, *supra* note 3.

13. *See* Samuelson, *Why Copyright Law Excludes*, *supra* note 3, at 1922.

14. 17 U.S.C. § 102(b); *see also* S. REP. NO. 93–983, at 107–08 (1974); H.R. REP. NO. 94–1476, at 56–57 (1976).

15. Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 363–64 (1991); *see also* Eldred v. Ashcroft, 537 U.S. 186, 219 (2003).

16. Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 548 (1985); *see also Eldred*, 537 U.S. at 219; Golan v. Holder, 565 U.S. 302, 328–29 (2012).

17. *See generally* RESTATEMENT OF COPYRIGHT § 12 cmt. e (AM. L. INST., Tentative Draft No. 2, 2022) (grouping unprotected elements into these three categories).

18. U.S. CONST. art. I, § 8, cl. 8.

unprotected categories.[19] This is not because unprotected elements are not valuable enough to justify copyright, but rather because they are so valuable that they belong in the public domain.[20] As Justice Brandeis famously put it: "The general rule of law is, that the noblest of human productions—knowledge, truths ascertained, conceptions, and ideas—become, after voluntary communication to others, free as the air to common use."[21] And as the Court later explained in *Feist Publications, Inc. v. Rural Telephone Service Co.*[22]: "[C]opyright assures authors the right to their original expression, but encourages others to build freely upon the ideas and information conveyed by a work. . . . This result is neither unfair nor unfortunate. It is the means by which copyright advances the progress of science and art."[23] Advancing the progress of knowledge requires allowing subsequent authors to craft their own expressive works that revisit existing ideas and to participate in the free exchange of facts.[24]

In addition to conforming copyright protection to the purpose embedded in the constitutional authorization, the idea/expression distinction is important for ensuring that copyright does not undermine patent law's related purpose of promoting progress in the useful arts.[25] Protection for methods, in particular, would put copyright law in tension with patent law. Patent protection is available for newly invented ways of doing things ("processes" in patent parlance), but only after a relatively rigorous examination to ensure that the invention is in fact novel and nonobvious.[26] It would disrupt this careful quid pro quo if the exclusive right to practice a method could be obtained based on the far less rigorous standards of copyright—which grants protection automatically to any work that is captured in a tangible medium and exhibits a

19. *See, e.g.*, Eichel v. Marcin, 241 F. 404, 408–09 (S.D.N.Y. 1913).

20. *See* Jessica Litman, *The Public Domain*, 39 EMORY L.J. 965, 967 (1990) (describing the public domain as "the law's primary safeguard of the raw material that makes authorship possible").

21. Int'l News Serv. v. Associated Press, 248 U.S. 215, 250 (1918) (Brandeis, J., dissenting).

22. 499 U.S. 340 (1991).

23. *Id.* at 349–50 (cleaned up).

24. *See* Sparaco v. Lawler, Matusky, Skelly, Eng'rs LLP, 303 F.3d 460, 466 (2d Cir. 2002) (describing how courts in the twentieth century developed the view "that historical, scientific, or factual information belongs in the public domain, and that allowing the first publisher to prevent others from copying such information would defeat the objectives of copyright by impeding rather than advancing the progress of knowledge"); Hoehling v. Universal City Studios, Inc., 618 F.2d 972, 978 (2d Cir. 1980) ("To avoid a chilling effect on authors who contemplate tackling an historical issue or event, broad latitude must be granted to subsequent authors who make use of historical subject matter, including theories or plots.").

25. *See* Peter S. Menell, *Rise of the API Copyright Dead?: An Updated Epitaph for Copyright Protection of Network and Functional Features of Computer Software*, 31 HARV. J.L. & TECH. 305, 418–20 (2018) [hereinafter Menell, *Rise of the API Copyright Dead?*] (describing how the law channels works among different modes of intellectual property protection and explaining that, otherwise, "[t]he long duration and low threshold requirements of copyright and trademark protection would displace patent's primacy in protecting technological advance or functional features").

26. 35 U.S.C. §§ 101–03.

mere "modicum" of creativity.[27] As the *Baker* Court explained: "To give to the author of the book an exclusive property in the art described therein, when no examination of its novelty has ever been officially made, would be a surprise and a fraud upon the public."[28]

The idea/expression distinction also helps to reconcile copyright law with the First Amendment, by ensuring that protection for a copyright owner's particular expression does not prohibit other people from speaking about the same topic.[29] The Supreme Court has referred to this as "a definitional balance between the First Amendment and the Copyright Act"[30] and as one of the "traditional contours" of copyright protection that serve as "built-in First Amendment accommodations."[31]

The upshot of the idea/expression distinction is that copying of unprotected elements alone does not constitute infringement of copyright in the work from which they are copied.[32] In other words, defendants cannot be held liable merely for *extracting* unprotected material, nor for making their own works deploying what they have extracted.[33]

Importantly, extraction and reuse of unprotectable elements from copyrighted works are permissible even if the extracted material provides the basis for a new work that competes in the marketplace with the original work. Cookbook authors may thus extract and replicate recipes from preexisting cookbooks even if the old and new cookbooks will compete in the same market.[34] Scholars may extract and replicate other scholars' scientific models from copyrighted notes (as a matter of copyright law, if not academic integrity), even if the scholars are competing for spots in prestigious publication outlets.[35] Yogis may extract and replicate other yogis' sequences of poses from copyrighted books, even for purposes of establishing competing yoga businesses.[36] Manufacturers of exercise machines may extract and replicate depictions of the

27. Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 346 (1991).

28. Baker v. Selden, 101 U.S. 99, 102 (1879). Keeping patent-like protection in its proper bounds also promotes competition and protects consumers from unreasonably high prices. *See* Samuelson, *Why Copyright Law Excludes*, *supra* note 3, at 1934.

29. *See* Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 582 (1985) (Brennen, J., dissenting).

30. *Id.* at 556 (majority opinion).

31. Eldred v. Ashcroft, 537 U.S. 186, 218–21 (2003); *see also* Golan v. Holder, 565 U.S. 302, 327–30 (2012).

32. Unprotected elements can be selected, arranged, or coordinated in original ways that qualify for copyright protection. In such cases, copying the original selection, arrangement, or coordination can amount to infringement. But merely extracting and copying the unprotected elements themselves cannot. *See, e.g.*, *Feist Publ'ns, Inc.*, 499 U.S. at 348.

33. *See* Samuelson, *Why Copyright Law Excludes*, *supra* note 3, at 1933.

34. *See* Publ'ns Int'l, Ltd. v. Meredith Corp., 88 F.3d 473, 482 (7th Cir. 1996).

35. *See* Seng-Tiong Ho v. Taflove, 648 F.3d 489, 498–500 (7th Cir. 2011).

36. Bikram's Yoga Coll. of India, L.P. v. Evolation Yoga, LLC, 803 F.3d 1032, 1044 (9th Cir. 2015).

body positions required to correctly perform exercises, even for purposes of competing with other machine producers.[37] Manufacturers of fasteners may extract and replicate parts numbers that correspond to a numbering system developed by another manufacturer, even for the purpose of more effectively attracting customers looking for fasteners with specific characteristics indicated by those numbers.[38] Numbers identifying transmission parts are similarly subject to extraction, in part to ensure that there is free competition in the market for (uncopyrighted) parts.[39] And roller derby promoters may copy the format of roller derby races from publications describing that format, even for the purposes of staging competing roller derby events![40] The freedom to extract reflects the importance of promoting, rather than restraining, competition on the basis of those elements to which copyright protection does not extend.[41]

In theory, extraction of unprotected elements from protected works could happen with surgical precision: ideas, methods, and facts could be plucked from an existing work and reused without fear of liability. In reality, there can be legal and practical impediments to extraction.

For one thing, there is no ex-ante process by which the protected and unprotected elements of copyrighted works are identified; they are identified definitively only in litigation. This means that a would-be extractor might accidentally copy and reuse some protected material along with unprotected material.

Imagine this hypothetical: a graphic artist is interested in creating a book cover that depicts a deceased scientist who was rarely photographed during her lifetime. The artist copies the scientist's likeness from a rare photo that appeared in a magazine when the scientist won the Nobel Prize. The artist's intention is only to copy the aspects of the photograph that represent the scientist's actual appearance. These facts are not themselves subject to copyright

37. Universal Athletic Sales Co. v. Salkeld, 511 F.2d 904, 909 (3d Cir. 1975).

38. Southco, Inc. v. Kanebridge Corp., 390 F.3d 276, 282–85 (3d Cir. 2004); Am. Dental Ass'n v. Delta Dental Plans Ass'n, 126 F.3d 977, 981 (7th Cir. 1997). *See generally* Pamela Samuelson, *Questioning Copyrights in Standards*, 48 B.C. L. REV. 193 (2007) (discussing line of cases involving parts numbers).

39. ATC Distrib. Grp., Inc. v. Whatever It Takes Transmissions & Parts, Inc., 402 F.3d 700, 709 (6th Cir. 2005).

40. Seltzer v. Sunbrock, 22 F. Supp. 621, 630 (S.D. Cal. 1938).

41. *See* Pamela Samuelson, *Functional Compilations*, 54 HOUS. L. REV. 321, 352 (2016) (observing, with regard to *Southco, Inc.*'s rejection of copyright protection for a parts numbering system, that "[l]urking in the background . . . was a danger that the claim of copyright in part numbering systems would give the plaintiffs a 'monopoly on the typically uncopyrightable product' to which the names and numbers were assigned" (quoting SmithKline Beecham Consumer Healthcare, L.P. v. Watson Pharm., Inc., 211 F.3d 21, 29 n.5 (2d Cir. 2000))).

protection.[42] But in the course of copying what the artist perceives to be the scientist's facial features, he inadvertently copies elements of the photo that resulted from the photographer's creative choices with regard to the interplay of light and shadow. Although it is extremely difficult for the artist to know in advance, it is possible that a court could later determine that those copied elements were protected expression, not unprotected facts.

Even where a would-be user is confident (and correct) that what they intend to copy is unprotected material, they may find it necessary to copy protected material in the act of extracting the unprotected elements. Computer software presents one context in which this dilemma is common. Courts have held that software interfaces that are necessary to make one computer program interoperate with another program, or with computer hardware, are unprotectable.[43] But it can be virtually impossible for a second programmer to identify (and thereby extract) the unprotected interfaces within a given program without first copying that entire program in a process known as decompilation.[44] Copying the entire program invariably involves copying protected expression, even if the ultimate purpose is to copy only the unprotected interfaces.

Both of these scenarios combine extraction of unprotected materials with incidental copying of protected materials (either by accident or necessity). I refer to this combination as "extractive use." Like "fair use" or "transformative use,"[45] this terminology refers to activity that may constitute prima facie

42. *See, e.g.*, Meshwerks, Inc. v. Toyota Motor Sales U.S.A., Inc., 528 F.3d 1258, 1264 (10th Cir. 2008) (explaining that "in the case of photographs, . . . authors are entitled to copyright protection only for the 'incremental contribution' . . . represented by their interpretation or expression of the objects of their attention" (quoting SHL Imaging, Inc. v. Artisan House, Inc., 117 F. Supp. 2d 301, 311 (S.D.N.Y. 2000))).

43. *See, e.g.*, Lexmark Int'l, Inc. v. Static Control Components, Inc., 387 F.3d 522, 535 (6th Cir. 2004), *abrogated on other grounds by* eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388 (2006); Comput. Assocs. Int'l, Inc. v. Altai, Inc., 982 F.2d 693, 709–10 (2d Cir. 1992); *see also* Peter S. Menell, *Envisioning Copyright Law's Digital Future*, 46 N.Y.L. SCH. L. REV. 63, 65–66 (2003) (describing evolution of the case law and observing that "[c]opyright law provides a thin layer of protection for computer software, effectively prohibiting wholesale piracy of computer programs without affording control for interface specifications and other essential elements of computer functionality" and "[t]he courts have also allowed subsequent software developers some leeway to reverse engineer software programs in order to develop interoperable programs"); Brief for Int'l Bus. Machs. Corp. and Red Hat, Inc. as Amici Curiae Supporting Petitioner, Google LLC v. Oracle Am., Inc., 593 U.S. 1 (2020) (No. 18-956) (arguing that unrestricted use of software interfaces is well-established and essential to software development and technological innovation generally).

44. *See* Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1515 (9th Cir. 1992); Dennis S. Karjala, *Copyright Protection of Computer Documents, Reverse Engineering, and Professor Miller*, 19 U. DAYTON L. REV. 975, 992 (1994).

45. *See* Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990) (developing the concept of transformative use and explaining that if a secondary use "adds value to the original—if the quoted matter is used as raw material, *transformed* in the creation of new information,

infringement of copyright (because expression is copied) but is not necessarily infringing in light of its purpose.[46]

Extractive use overlaps to some extent with related terminology that scholars and judges have used to describe uses of copyrighted works that are not designed merely to exploit the expressive value of those works, including "non-consumptive use" and "non-expressive use."[47] As Matthew Sag has described, "[t]he term 'non-expressive use' (also referred to by some as 'non-consumptive use') refers to any act of reproduction that is not intended to enable human enjoyment, appreciation, or comprehension of the copied expression as expression."[48] Or, as James Grimmelmann puts it: "[N]onexpressive uses do not count as reading. They are not part of the market that copyright cares about, because the author's market consists only of readers."[49] In their work on "fair learning," Mark Lemley and Brian Casey argue that "[c]opyright law should permit copying of works for non-expressive purposes—at least in most circumstances," and count learning about unprotected "ideas, facts, or functions" as one such purpose.[50] BJ Ard augments these concepts by focusing

new aesthetics, new insights and understandings—this is the very type of activity that the fair use doctrine intents to protect for the enrichment of society") (emphasis added). See *infra* Section II.B.1 for a discussion of the role of "transformative use" in the fair use analysis.

46. *But see* Oren Bracha, *The Work of Copyright in the Age of Machine Production*, HARV. J.L. & TECH. (forthcoming 2025), https://ssrn.com/abstract=4581738 [https://perma.cc/9Y9A-QFXV] [hereinafter Bracha, *Work of Copyright*] (arguing that neither reproduction of copyrighted works strictly for the purpose of training generative artificial intelligence systems nor producing new works that replicate the style of preexisting works amounts to prima facie copyright infringement).

47. *See, e.g.*, Sag, *The New Legal Landscape*, *supra* note 3, at 301 (defining "non-expressive use"); Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. 1607, 1624–57 (2009) [hereinafter Sag, *Copyright and Copy-Reliant Technology*] (explaining the doctrinal implications of "non-expressive use"); Lemley & Casey, *supra* note 3, at 750 ("Copyright law should permit copying of works for non-expressive purposes—at least in most circumstances."); Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 51–57 (2017) (describing non-expressive fair use); Matthew Sag, *Orphan Works as Grist for the Data Mill*, 27 BERKELEY TECH. L.J. 1503, 1512–42 (2012) [hereinafter Sag, *Orphan Works*] (discussing non-expressive use and arguing that it should generally be favored in fair use analysis); Matthew Sag, *The Google Book Settlement and the Fair Use Counterfactual*, 55 N.Y.L. SCH. L. REV. 19, 54 (2010) (describing "non-consumptive research"); James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 661–65 (2016) [hereinafter Grimmelmann, *Literate Robots*] (describing "non-expressive reading"); Amended Settlement Agreement at § 1.93, Authors Guild, Inc. v. Google, Inc., Case No. 05 CV 8136-DC (S.D.N.Y. Nov. 13, 2009) (defining non-consumptive research as "research in which computational analysis is performed on one or more Books, but not research in which a researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book"). *See generally* Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful*, 53 U.C. DAVIS L. REV. 893, 937 (2019) (discussing the applicability and limits of the "non-expressive" and "non-consumptive" terminology as applied to text data mining).

48. Sag, *The New Legal Landscape*, *supra* note 3, at 301.

49. Grimmelmann, *Literate Robots*, *supra* note 47, at 664.

50. Lemley & Casey, *supra* note 3, at 750.

on uses that exploit only the "non-authorial value" of preexisting works; Michael Mattioli adds the concept of "facilitative fair use."[51]

Many uses that I would classify as extractive uses could also be fairly classified as "non-expressive" or "non-consumptive." And many uses that could be classified as "non-expressive" or "non-consumptive" are also extractive. The objects of extractive use in my formulation are also the very elements that Ard identifies as reflecting "non-authorial value." But the concept of extractive use differs from these concepts (and aligns more closely with Casey and Lemley's concept of "fair learning" and Mattioli's theory of "facilitative fair use") in its focus on what the user *is* doing (extracting unprotected elements) as opposed to what the user is *not* doing (consuming the work's expression or exploiting its authorial value). This focus on the affirmative essence of extractive use helps to explain why extractive use should be permissible even when it also involves some consumption of expressive or authorial elements of existing copyrighted works, as in the scenario of the artist described above.

In sum, the idea/expression distinction is core to copyright's successful promotion of intellectual progress, its fostering of free expression, and its compatibility with patent law. The distinction can be understood as a negative limit on copyright protection, and also as an affirmative freedom to extract those elements of copyrighted works to which protection does not extend. Full validation of the freedom to extract requires permitting "extractive use" of copyrighted works, even where it involves incidental use of expression. In the next section, I explore how courts have upheld the freedom to extract by permitting extractive use—noting along the way some instances in which courts have fallen short.

B. *Doctrinal Recognition of the Freedom to Extract*

The idea/expression distinction and its statutory manifestation in 17 U.S.C. § 102(b) provide the doctrinal foundation for the freedom to extract. But, as noted above, the denial of copyright protection for ideas, methods, and facts is not always sufficient to protect the freedom to extract as a practical matter. Legal and logistical impediments to perfect extraction mean that extractive use of some expression can be necessary to fully vindicate this freedom. A number of doctrines in copyright law recognize and address this dilemma. This section describes those doctrines, their strengths, and their shortcomings. The doctrines include merger, scenes à faire, fair use, and the useful article doctrine.

51. BJ Ard, *Copyright's Latent Space: Generative AI and the Limits of Fair Use*, 110 CORNELL L. REV. (forthcoming 2025) (manuscript at 51) (on file with the North Carolina Law Review); Michael Mattioli, *Facilitative Fair Use*, 102 DENV. L. REV. (forthcoming 2025) (on file with the North Carolina Law Review).

1. Merger

The doctrine that most directly addresses the practical difficulties of extracting unprotected elements from copyrighted works is merger. The merger doctrine recognizes that some unprotected ideas, methods, or facts do not lend themselves to a wide range of expression.[52] This means that giving exclusive rights based on one author's expression of such unprotected material could give the copyright owner effective control over the unprotected element.[53] A widely cited case illustrates this dynamic. In *Morrissey v. Procter & Gamble Co.*,[54] the plaintiff claimed that the defendant had infringed the plaintiff's copyright in the rules of a promotional sweepstakes.[55] The First Circuit held, citing *Baker v. Selden*, that "the *substance* of the contest was not copyrightable."[56] That is, the plaintiff's copyright in the text of the sweepstakes rules could not prevent others from operating a sweepstakes that worked in the same way—just as the publication of a book describing a system of accounting could not, in *Baker*, prevent others from practicing that system.[57] But that still left the question of whether the defendant had infringed the copyright in the text of the sweepstakes rules by using text that demonstrated "almost precise similarity" with the plaintiff's.[58] In the course of extracting the substance of the sweepstakes, had the defendant infringed the copyright in the words used to express that substance? The court said no, recognizing that protecting the expression under such circumstances would have the effect of insulating the substance from lawful extraction:

> When the uncopyrightable subject matter is very narrow, so that the topic necessarily requires, if not only one form of expression, at best only a limited number, to permit copyrighting would mean that a party or parties, by copyrighting a mere handful of forms, could exhaust all possibilities of future use of the substance. In such circumstances it does not seem accurate to say that any particular form of expression comes from the subject matter. However, it is necessary to say that the subject matter would be appropriated by permitting the copyrighting of its

52. *See* RESTATEMENT OF COPYRIGHT § 17, cmt. b (AM. L. INST., Tentative Draft No. 5, 2024) ("Merger typically applies when an unprotectable element can be expressed in only one way, or very few ways; the key inquiry is not exactly how many alternatives there are, but rather whether the range of viable alternatives is so constrained that granting copyright would effectively bar others from expressing the unprotected element.").

53. *Id.*

54. 379 F.2d 675 (1st Cir. 1967).

55. *Id.* at 675–76.

56. *Id.* at 678 (emphasis added).

57. *Id.*; Baker v. Selden, 101 U.S. 99, 107 (1879).

58. *Morrissey*, 379 F.2d at 678.

> expression. We cannot recognize copyright as a game of chess in which the public can be checkmated.[59]

Baker v. Selden can be understood to support this conclusion as well. *Baker* held not only that the plaintiff's copyright did not give exclusive rights to the bookkeeping system, but also that it did not forbid use of forms that were "necessary incidents" to that system:

> [W]here the art [a book] teaches cannot be used without employing the methods and diagrams used to illustrate the book, or such as are similar to them, such methods and diagrams are to be considered as necessary incidents to the art, and given therewith to the public; not given for the purpose of publication in other works explanatory of the art, but for the purpose of practical application.[60]

The denial of protection to seemingly expressive elements that represent the only (or one of only a few) ways to express an idea, system, or fact has come to be known as the "merger" doctrine, although neither *Morrissey* nor *Baker* used that term. Courts applying the merger doctrine generally recognize that, without it, the freedom to extract unprotected elements would be illusory.[61] Every attempt by a subsequent author to extract and then re-express the unprotected elements would trigger liability because the resulting work would necessarily be the same as (or at least substantially similar to) the preexisting work from which those elements had been extracted.

In only rare cases does the merger doctrine render an entire work uncopyrightable. More frequently, only some expressive elements within a copyrightable work are held to be unprotected, thus limiting the scope of protection in that work. This limited scope typically plays out in the context of the infringement analysis. Various elements of the allegedly infringed work are identified, unprotected elements are eliminated from consideration, and the remaining elements are compared to what the defendant copied to determine whether that copying amounted to improper appropriation of protected elements. The steps in this process were described in the context of alleged

59. *Id.* at 678–79 (cleaned up).

60. *Baker*, 101 U.S. at 103.

61. *See, e.g.*, Herbert Rosenthal Jewelry Corp. v. Kalpakian, 446 F.2d 738, 742 (9th Cir. 1971) (explaining that "[w]hen the 'idea' and its 'expression' are thus inseparable, copying the 'expression' will not be barred, since protecting the 'expression' in such circumstances would confer a monopoly of the 'idea' upon the copyright owner free of the conditions and limitations imposed by the patent law"); Design Basics, LLC v. Signature Constr., Inc., 994 F.3d 879, 889 (7th Cir. 2021) ("Merger doctrine prevents the use of copyright to protect an idea or procedure."); Kay Berry, Inc. v. Taylor Gifts, Inc., 421 F.3d 199, 209 (3d Cir. 2005) ("In some instances, there may come a point when an author's expression becomes indistinguishable from the idea he seeks to convey, such that the two merge. In these circumstances, no protection is available for the expression; otherwise, the copyright owner could effectively acquire a monopoly on the underlying art or the idea itself." (citation omitted)).

infringement of computer software in *Computer Associates International, Inc. v. Altai, Inc.*,[62] where the Second Circuit articulated an approach in which "a court would first break down the allegedly infringed program into its constituent structural parts."[63] Next, "by examining each of these parts for such things as incorporated ideas, expression that is necessarily incidental to those ideas, and elements that are taken from the public domain, a court would then be able to sift out all non-protectable material" and identify "a kernel, or possible kernels, of creative expression."[64] Finally, the court would compare this expression with the allegedly infringing program to determine whether there was sufficient similarity to support a finding of infringement.[65] The *Altai* court went on to identify merger as one doctrine that helps to identify elements that should be filtered out in the second step of what has come to be known as the abstraction-filtration-comparison approach to copyright infringement.[66]

Most courts understand the merger doctrine to apply even if there are multiple ways to express the unprotected element(s) at issue, so long as the range of options is so narrow that protection of the expression would, as a practical matter, extend to the unprotected material.[67] A few courts, by contrast, only apply merger to deny copyrightability when there is exactly one way to express the unprotected elements at issue.[68] But these courts also acknowledge the threat to the freedom to extract that is posed by a narrow range of expressive options. In such circumstances, these courts hold that the expression is copyrightable, but that copying it will be actionable only where the allegedly infringing work is "virtually identical," not merely substantially similar, to the preexisting work.[69] Under either approach, the merger doctrine aims to overcome practical obstacles to exercise of the freedom to extract.

2. Scènes à Faire

Closely related to merger, the scènes à faire doctrine also helps to safeguard the freedom to extract by permitting the copying of expression that naturally accompanies unprotected subject matter. Under this doctrine, courts

62. 982 F.2d 693 (2d Cir. 1992).

63. *Id.* at 706.

64. *Id.*

65. *Id.*

66. *Id.* at 708.

67. *See, e.g.*, N.Y. Mercantile Exch., Inc. v. Intercontinental Exch., Inc., 497 F.3d 109, 117 n.9 (2d Cir. 2007); Design Basics, LLC v. Signature Constr., Inc., 994 F.3d 879, 889–90 (7th Cir. 2021). For an explanation of why merger may be found even if there is more than one way to express an idea, see Pamela Samuelson, *Reconceptualizing Copyright's Merger Doctrine*, 63 J. COPYRIGHT SOC'Y U.S.A. 417, 425–28 (2016) [hereinafter Samuelson, *Reconceptualizing Copyright's Merger Doctrine*].

68. Ets-Hokin v. Skyy Spirits, Inc., 225 F.3d 1068, 1082 (9th Cir. 2000).

69. *See, e.g.*, Yankee Candle Co. v. Bridgewater Candle Co., LLC, 259 F.3d 25, 36 (1st Cir. 2001); Apple Comput., Inc. v. Microsoft Corp., 35 F.3d 1435, 1439 (9th Cir. 1994); *see also* Ets-Hokin v. Skyy Spirits, Inc., 323 F.3d 763, 766 (9th Cir. 2003).

deny protection "to those expressions that are standard, stock, or common to a particular topic or that necessarily follow from a common theme or setting."[70] As with the merger doctrine and the idea/expression distinction more generally, courts deploy scènes à faire to ensure that copyright protection does not yield exclusive control over the fundamental building blocks of knowledge and creativity.[71] And so, movies about superheroes may feature descriptions of the heroes' ability to fly, their secret double lives, and their megalomaniacal enemies, even though those details have appeared in superhero movies before (and even if a defendant copied those elements from preexisting examples).[72] Once such features have become standard within a genre, their inclusion can become part-and-parcel of engaging with that genre and its recurring ideas.

One way to understand the scènes à faire doctrine is that it allows subsequent authors to communicate about unprotectable ideas using expression that taps into established mental and emotional connections with audiences who have previously been exposed to those ideas—that is, to communicate in a way that is compatible with audience expectations.[73] Similarly, scènes à faire has been invoked to ensure that subsequent computer programmers can connect to interfaces in a way that allows compatibility with existing hardware and software (and the user base that has invested in that hardware and software). As the Sixth Circuit has explained, "[i]n the computer-software context, the doctrine means that the elements of a program dictated by practical realities—e.g., by hardware standards and mechanical specifications, software standards and compatibility requirements, computer manufacturer design standards, target industry practices, and standard computer programming practices—may not obtain protection."[74]

These compatibility concerns come to the fore when manufacturers claim copyright protection for what are often referred to as "lock-out codes"—computer programs that perform an authentication function necessary for

70. Gates Rubber Co. v. Bando Chem. Indus., Ltd., 9 F.3d 823, 838 (10th Cir. 1993).

71. *See, e.g.*, *Ets-Hokin*, 225 F.3d at 1082 (explaining that the rationale for the scènes à faire doctrine, "like merger, . . . is that there should be no monopoly on the underlying unprotectable idea"); *Gates Rubber Co.*, 9 F.3d at 838 (explaining that "[u]nder the *scenes a faire* doctrine, we deny protection to those expressions that are standard, stock, or common to a particular topic or that necessarily follow from a common theme or setting," because "[g]ranting copyright protection to the necessary incidents of an idea would effectively afford a monopoly to the first programmer to express those ideas").

72. *See* Lewinson v. Henry Holt & Co., LLC, 659 F. Supp. 2d 547, 567–68 (S.D.N.Y. 2009) (providing examples of unprotected scènes à faire elements).

73. *See generally* Bo S.L. Kim, *Copyright's Public Reliance Interests*, 99 WASH. L. REV. 107 (2024) (developing a theory of the scènes à faire doctrine focused on audience expectations and public reliance interests).

74. Lexmark Int'l, Inc. v. Static Control Components, Inc., 387 F.3d 522, 535 (6th Cir. 2004), *abrogated on other grounds by* eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388 (2006).

interoperability.[75] A competitor might copy a lock-out code for the purpose of extraction—that is, to practice the unprotectable method that allows two components to interoperate. Plaintiffs have nonetheless claimed that their lock-out codes contain expressive elements that may not be copied without authorization.[76] Courts have generally rejected those arguments and thereby vindicated the freedom to extract, often deploying the scènes à faire or merger doctrines.

For example, in *Lexmark International, Inc. v. Static Control Components, Inc.*,[77] the Sixth Circuit considered an infringement claim by a hardware manufacturer attempting to constrain the aftermarket components that could be used with its equipment.[78] Plaintiff Lexmark manufactures printers and toner cartridges. At the time of the suit, Lexmark had begun selling discount toner cartridges that contained a microchip designed to prevent the cartridges from functioning with Lexmark printers if the cartridges had been refilled by an unauthorized remanufacturer.[79] Lexmark sued defendant Static Control Components, a company that sells computer chips used in third-party remanufactured toner cartridges.[80] Lexmark alleged, inter alia, that Static Control infringed Lexmark's copyright by wholesale copying of the "toner loading program" embedded in Lexmark's toner cartridges.[81] The Sixth Circuit rejected that argument, describing the toner loading program as a lock-out code and explaining that "[t]o the extent compatibility requires that a particular code sequence be included in the component device to permit its use, the merger and scènes à faire doctrines generally preclude the code sequence from obtaining copyright protection."[82]

Other courts have similarly held that lock-out codes, authentication sequences, communications protocols, and other elements essential to the operation of multi-component systems are unprotectable as ideas, methods of operation, merged expression, or scènes à faire.[83] Together, these doctrines help to ensure that neither the requirements of a particular genre, nor those of a

75. Regarding lock-out codes and the economics of network effects, see Peter S. Menell, *Economic Analysis of Network Effects and Intellectual Property*, *in* 1 RESEARCH HANDBOOK ON THE ECONOMICS OF INTELLECTUAL PROPERTY LAW 157, 179 (Ben Depoorter, Peter S. Menell & David Schwartz eds., 2019) [hereinafter Menell, *Economic Analysis of Network Effects*].

76. *See Lexmark*, 387 F.3d at 542.

77. *Id.*

78. *Id.* at 530.

79. *Id.*

80. *Id.* at 530–31.

81. *Id.* at 531.

82. *Id.* at 536. The court ultimately reversed the district court's grant of a preliminary injunction, holding that Lexmark was not likely to succeed on its infringement claim because the toner loading program operated as an uncopyrightable lock-out code. *Id.* at 536–44.

83. *See, e.g.*, Pyrotechnics Mgmt., Inc. v. XFX Pyrotechnics LLC, 38 F.4th 331, 339, 339 n.7 (3d Cir. 2022); Comput. Assoc. Int'l, Inc. v. Altai, Inc., 982 F.2d 693, 709 (2d Cir. 1992).

particular technological environment, serve as practical impediments to exercising the freedom to extract.

3. The Limits of Merger and Scènes à Faire

Both merger and scènes à faire tend to operate at the margins, forgiving defendants from copying a bit more expression than would otherwise be permitted in order to safeguard the freedom to extract. Both doctrines have flexibility to adjust to technological change. For example, both doctrines have been applied in ways that acknowledge how software and hardware specifications can constrain the expressive options available to computer programmers. But neither doctrine has proven sufficient to address a different technological development, to which I now turn.

Sometimes the obstacle to exercising the freedom to extract is that the unprotected elements within a work *cannot even be identified* for purposes of extracting them. This dilemma did not exist when courts first articulated the idea/expression distinction and the merger and scènes à faire doctrines connected to it. Copyright protects works that are fixed in a tangible medium of expression from which they can be perceived. This generally means that people can see or hear a copyrighted work, and thus have access to the ideas, methods, and facts that the work contributes to the public domain.[84] But copyright protection for computer software, expressly recognized since the 1976 Copyright Act, changes that. Software is typically distributed in a form—object code—that performs the software's functions without necessarily communicating to human beings the ideas, methods, or facts that make those functions possible.[85] Software object code is readable by the computers that operate the software but not by humans who might want to understand it.

In theory, the idea/expression distinction, merger, and scènes à faire combine to allow subsequent programmers to reuse some of the basic building blocks embedded in copyrighted software. In practice, those building blocks can be virtually impossible to discover without using techniques that entail copying the software (including its expressive elements) in its entirety. Specifically, programmers interested in extracting from an existing program only unprotected elements—for example, those elements necessary to allow interoperability with a new program, or with different hardware—often need to copy the existing program as part of the reverse engineering process of

84. *See* Karjala, *supra* note 44, at 994 (contrasting computer programs with "[a]ll other copyright-protected works," which "carry their ideas and other unprotected elements on their face").

85. *See generally* Grimmelmann, *Literate Robots*, *supra* note 47 (contrasting copyright law's treatment of human versus "robotic" reading); Karjala, *supra* note 44, at 994 ("Computer programs in object-code form are alone among all types of publicly distributed copyright-protected works in being unreadable by human beings.").

"disassembly" or "decompilation."[86] Copying the entire program for this purpose invariably involves copying protected expression, even if the ultimate purpose is to copy only the unprotected interfaces.

Courts have repeatedly recognized that interfaces that allow programs to interoperate with each other, or with specific hardware platforms, or even with the acquired knowledge and expectations of consumers, should fall on the unprotected side of the idea/expression distinction. If those interfaces require a specific form of expression, the expression should be unprotected on the basis of merger or scènes à faire.[87] But the idea/expression distinction, even as augmented by merger and scènes à faire, cannot easily be stretched to excuse the verbatim copying of an entire work. And yet, that copying can be a necessary step to extracting those unprotected elements that the idea/expression, merger, and scènes à faire doctrines purport to liberate.[88] As the next section explains, courts addressing this scenario have generally turned to the more flexible fair use doctrine to address the question of intermediate copying conducted for purposes of extraction.

4. Fair Use

The fair use doctrine is a context-specific rule of reason that serves many purposes within copyright law. Among its other functions, it can safeguard the freedom to extract. Courts have deployed fair use this way most frequently in the context of computer software. This is a context, as just described, in which users often find it necessary to copy protected material in the act of extracting unprotected elements.

Sometimes courts deploy fair use under circumstances in which the idea/expression distinction, merger, or scènes à faire might independently have permitted the copying in question. In *Google LLC v. Oracle America, Inc.*,[89] for example, the Supreme Court held that Google's copying of Oracle's application programming interfaces was fair use without grappling with the question of whether those interfaces contained protectable expression.[90] The fair use holding ultimately vindicated the freedom to extract those interfaces, but one could imagine the same result from careful application of the scope-limiting

86. For the Ninth Circuit's description of this process, see Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1515 (9th Cir. 1992); *see also* Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596, 599 (9th Cir. 2000).

87. *See, e.g.*, *Lexmark*, 387 F.3d at 535.

88. *See* Jerome H. Reichman, Graeme B. Dinwoodie & Pamela Samuelson, *A Reverse Notice and Takedown Regime to Enable Public Interest Uses of Technically Protected Copyrighted Works*, 22 BERKELEY TECH. L.J. 981, 1030–31 (2007) (assessing the "broader implications" of the lock-out technology cases).

89. 593 U.S. 1 (2021).

90. *Id.* at 20.

doctrines described above.[91] Indeed, the district court had originally decided the case on the basis that the elements Google copied were not copyrightable.[92]

In other circumstances involving the freedom to extract, by contrast, fair use has played a more indispensable role. The following sections illustrate this across a range of factual scenarios, making frequent reference to the four fair use factors the Copyright Act directs courts to consider: (1) the purpose and character of the use; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion of the copyrighted work used; and (4) the effect of the use on the potential market for or value of the copyrighted work.[93]

a. Reverse Engineering Software Object Code

Copying expression can be necessary to exercise the freedom to extract by means of reverse engineering copyrighted computer software.[94] For example, in *Sega Enterprises Ltd. v. Accolade, Inc.*,[95] the Ninth Circuit held that the defendant's copying of plaintiff's software for purposes of reverse engineering its interfaces to produce competing games for plaintiff's video game console was fair use.[96] The court summarized its conclusion in terms that resonate strongly with the freedom to extract:

> Although the question is fairly debatable, we conclude based on the policies underlying the Copyright Act that disassembly of copyrighted object code is, as a matter of law, a fair use of the copyrighted work if such disassembly provides the only means of access to those elements of the code that are not protected by copyright and the copier has a legitimate reason for seeking such access.[97]

The fact that the defendant in *Sega* was extracting unprotected elements contributed to the court's analysis of all four of the fair use factors specified by the Copyright Act.[98] First, considering the "purpose and character of the use,"[99] the court held that defendant Accolade's purpose, "to discover the functional requirements for compatibility with the Genesis console—aspects of Sega's

91. *See* Google's Opening Copyright Liability Trial Brief at 4–10, Oracle Am., Inc. v. Google, Inc., 872 F. Supp. 2d 974 (N.D. Cal. 2012) (No. 3:10-CV-03561-WHA).

92. *Oracle*, 872 F. Supp.at 1002.

93. 17 U.S.C. § 107.

94. *See generally* Pamela Samuelson & Suzanne Scotchmer, *The Law and Economics of Reverse Engineering*, 111 YALE L.J. 1575 (2002) (analyzing the legal status of reverse engineering); Menell, *Rise of the API Copyright Dead?*, *supra* note 25 (analyzing API copyright disputes).

95. 977 F.2d 1510 (9th Cir. 1992).

96. *Id.* at 1518.

97. *Id.*; *see also* Bateman v. Mnemonics, Inc., 79 F.3d 1532, 1539 n.18 (11th Cir. 1996).

98. *Sega*, 977 F.2d at 1520.

99. 17 U.S.C. § 107(1).

programs that are not protected by copyright" was "a legitimate, essentially non-exploitative purpose."[100]

On the second fair use factor, "the nature of the copyrighted work,"[101] the court acknowledged that the program at issue—while largely functional—contained both protected and unprotected elements and that Accolade had copied all of them as part of the disassembly process.[102] But the critical aspect of the nature of the work was that it was the type of work for which this disassembly was necessary in order to extract unprotected elements.[103] The court explained that this dictated the analysis of the second fair use factor in favor of fair use[104]: "Because Sega's video game programs contain unprotected aspects *that cannot be examined without copying*, we afford them a lower degree of protection than more traditional literary works."[105] The court understood that where the nature of the copyrighted work in question is software distributed in object code, the freedom to extract unprotected elements would be meaningless as a practical matter without the possibility of copying protected elements as a necessary step in the extraction process.[106] Weighing this factor in favor of fair use helps to avoid that result.

As to the third fair use factor, "the amount and substantiality of the portion used in relation to the copyrighted work as a whole," the court observed that the defendant had copied the plaintiff's work in its entirety and held that this factor therefore counted in the plaintiff's favor.[107] But "where the ultimate (as opposed to direct) use is as limited as it was here, the factor is of very little weight."[108] Again, the defendant's ultimate purpose—extracting unprotected elements in order to generate new works that were not substantially similar to the plaintiff's—shaped the court's analysis of this factor.[109] Indeed, one could

100. *Sega*, 977 F.2d at 1522–23; *see also* Lexmark Int'l, Inc. v. Static Control Components, Inc., 387 F.3d 522, 544 (6th Cir. 2004), *abrogated on other grounds by* eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388 (2006) (distinguishing, for purposes of the first factor, between use of a program to permit hardware functionality versus using it "for its commercial value *as a copyrighted work*").

101. 17 U.S.C. § 107(2).

102. *Sega*, 977 F.2d at 1525.

103. *Id.*

104. *Id.* at 1526.

105. *Id.* (emphasis added).

106. *See* Atari Games Corp. v. Nintendo of Am., Inc., 975 F.2d 832, 843 (Fed. Cir. 1992) ("When the nature of a work requires intermediate copying to understand the ideas and processes in a copyrighted work, that nature supports a fair use for intermediate copying. Thus, reverse engineering object code to discern the unprotectable ideas in a computer program is a fair use."). Regarding the relationship between the second factor and the defendant's purpose to extract unprotected information about the copyrighted work, see Authors Guild v. Google, Inc., 804 F.3d 202, 220 (2d Cir. 2015) (suggesting that the second factor can favor fair use where "the secondary use transformatively provides valuable information about the original, rather than replicating protected expression in a manner that provides a meaningful substitute for the original").

107. *Sega*, 977 F.2d at 1526.

108. *Id.* at 1526–27.

109. *Id.* at 1527.

imagine these facts leading a court to conclude that complete copying would not necessarily put the third factor in the plaintiff's column at all.

With regard to the fourth fair use factor, "the effect of the use upon the potential market for or value of the copyrighted work," the court again focused on safeguarding the freedom to engage in activity that copyright law does not forbid—namely, competing in the market for a *type* of work by introducing new works that are not substantially similar to existing works.[110] This type of legitimate competition on the basis of unprotected elements should not count as an effect on the market that weighs against fair use. Indeed, attempting to squash that type of competition runs "counter to the statutory purpose of promoting creative expression and cannot constitute a strong equitable basis for resisting the invocation of the fair use doctrine."[111] The court therefore concluded that the fourth factor counted in defendant Accolade's favor.[112]

The *Sega* court summed up its fair use analysis by again emphasizing how the nature of computer software means that copyright—if not deployed with sensitivity to the subject matter—could threaten the freedom to extract:

> [T]he fact that computer programs are distributed for public use in object code form often precludes public access to the ideas and functional concepts contained in those programs, and thus confers on the copyright owner a *de facto* monopoly over those ideas and functional concepts. That result defeats the fundamental purpose of the Copyright Act—to encourage the production of original works by protecting the expressive elements of those works while leaving the ideas, facts, and functional concepts in the public domain for others to build on.[113]

The Ninth Circuit revisited these themes in *Sony Computer Entertainment Inc. v. Connectix Corp.*,[114] in which defendant Connectix had copied plaintiff Sony's video game console firmware (the "basic input-output system," or "BIOS") in the course of reverse engineering it to develop emulator software that would allow the plaintiff's video games to be played on regular multipurpose computers.[115] The Ninth Circuit reversed the district court's grant of a preliminary injunction, holding that "[t]he intermediate copies made and used by Connectix during the course of its reverse engineering of the Sony BIOS were protected fair use, necessary to permit Connectix to make its non-

110. *Id.* at 1522–24.

111. *Id.* at 1523–24.

112. *Id.* at 1524.

113. *Id.* at 1527; *see also* Grimmelmann, *Literate Robots*, *supra* note 47, at 662 ("[T]he conceptual twist in *Sega v. Accolade* is crucial, because it stands for the principle that non-expressive reading does not count as infringement. That principle is much broader than software; it applies whenever there is something to be learned about a copyrighted work other than its expressive authorship.").

114. 203 F.3d 596 (9th Cir. 2000).

115. *Id.* at 598, 601.

infringing Virtual Game Station function with PlayStation games."[116] Again, the court described its conclusion in terms that capture the essence of the freedom to extract. It explained that the "unprotected ideas and functions of [object] code . . . are frequently undiscoverable in the absence of investigation and translation that may require copying the copyrighted material."[117] Copying is necessary in this context to exercise the freedom to extract, and therefore "Connectix's intermediate copying and use of Sony's copyrighted BIOS was a fair use for the purpose of gaining access to the unprotected elements of Sony's software."[118]

The Ninth Circuit's consideration of the four fair use factors in *Sony* echoed its analysis in *Sega*. For example, it cited *Sega* in its analysis of the fourth factor, concluding that even if Connectix's Virtual Game Station emulator competed in the marketplace with Sony's PlayStation, that would not "compel a finding of no fair use."[119] Instead, the court viewed Connectix's Virtual Game Station (which did not itself contain any of Sony's copyrighted material) as a "legitimate competitor in the market for platforms on which Sony and Sony-licensed games can be played."[120] The court acknowledged, but dismissed, the fact that Sony might lose sales and profits from this competition: "Sony understandably seeks control over the market for devices that play games Sony produces or licenses. The copyright law, however, does not confer such a monopoly."[121]

The Federal Circuit's analysis in *Atari Games Corp. v. Nintendo of America, Inc.*[122] also resonates strongly with the freedom to extract, illustrating why the flexibility of fair use may be necessary to safeguard that right in response to technological developments that might obscure a work's unprotectable elements. In *Atari*, the court explained that copyright should not facilitate the obfuscation of the elements it dedicates to the public domain (or else to the more demanding realm of patent protection): "An author cannot acquire patent-like protection by putting an idea, process, or method of operation in an unintelligible format and asserting copyright infringement against those who try to understand that idea, process, or method of operation."[123] The court identified fair use as the doctrine through which copyright law "permits an individual in rightful possession of a copy of a work to undertake necessary efforts to understand the work's ideas, processes, and methods of operation."[124]

116. *Id.* at 599.
117. *Id.* at 602.
118. *Id.* (citations omitted); *see also id.* at 603.
119. *Id.* at 607 (citing Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1522–23 (9th Cir. 1992)).
120. *Id.* (citing *Sega*, 977 F.2d at 1522–23).
121. *Id.* (citing *Sega*, 977 F.2d at 1523–24).
122. 975 F.2d 832 (Fed. Cir. 1992).
123. *Id.* at 842.
124. *Id.*

The freedom to extract must be a practical reality in order for it to function in service of copyright's ultimate goals. As this section has demonstrated, fair use is one doctrine that can adjust copyright to vindicate the freedom to extract notwithstanding practical constraints.

b. Extracting Data Embedded in Copyrightable Software and Databases

Object code is not the only format that can necessitate copying of expression in order to extract unprotectable elements. In *Assessment Technologies of WI, LLC v. WIREdata, Inc.*,[125] the Seventh Circuit considered a situation in which the owner of copyright in a computer program used by municipal employees to organize data argued that extraction of the data from that program (called Market Drive) would infringe the owner's copyright.[126] After explaining that the copyright on the program did not extend to the unprotectable data the program arranged, the Seventh Circuit went on to say that the plaintiff "would lose this copyright case even if the raw data were so entangled with Market Drive that they could not be extracted without making a copy of the program."[127] The case would then be governed by the fair use logic of *Sega Enterprises Ltd. v. Accolade, Inc.*[128] Similarly, it would be lawful for the defendant to copy the protected compilation of data (not merely the unprotected data themselves) where "the only purpose of the copying would be to *extract noncopyrighted material*."[129]

Even if it were not impossible, but merely more expensive, to extract the data without copying protectable elements of the plaintiff's work, the court held that the would-be extractor should not have to bear that expense: "[S]ince [the plaintiff] has no ownership or other legal interest in the data . . . , it has no legal ground for making the acquisition of that data more costly for [the defendant]. [The plaintiff] is trying to use its copyright to sequester uncopyrightable data . . . ."[130]

Several courts have followed *WIREdata* in holding that copyrighted software may be copied in order to extract unprotected data. For example, in *Evolution, Inc. v. SunTrust Bank*,[131] the court held that a customer who was dissatisfied with the software it had obtained under license could copy portions of that software in order to extract the defendant's own data in the process of

125. 350 F.3d 640 (7th Cir. 2003).

126. *Id.* at 643.

127. *Id.* at 644.

128. *Id.* at 644–45 (citing Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1520–28 (9th Cir. 1992)).

129. *Id.* at 645 (emphasis added).

130. *Id.*

131. 342 F. Supp. 2d 943 (D. Kan. 2004).

switching to a new software provider, concluding that "[s]uch use of plaintiff's source code falls well within the fair use doctrine . . . ."[132]

Other courts have distinguished *WIREdata* where the defendant copied the plaintiff's work not merely to extract unprotected elements, but to create something substantially similar to and competitive with the plaintiff's work.[133] This type of copying cannot be characterized as fully extractive use and does not have the same claim to fairness as the copying done merely to liberate unprotected elements from the technological formats in which they are embedded.

c. Text Data Mining

Another technological development has raised the stakes for the freedom to extract in recent decades. Emerging computational techniques and capacity have presented new opportunities for deploying the freedom to extract to generate and disseminate new and socially valuable knowledge. In some ways, these techniques are akin to old-fashioned extraction—they involve learning and deploying unprotected ideas, methods, and facts that are evident on the face of existing works. In traditional extraction situations, this process typically involves a human being reading (or viewing or listening) in order to distill information from an existing work. The processing of information in this way does not implicate copyright at all. And if the reader reuses the information they learned, for example, by producing a work of their own, that extraction is protected by the idea/expression distinction and the adjacent doctrines of merger and scènes à faire.

In the contemporary environment, this extraction process has been superpowered by data science. Vast bodies of existing works can be processed by computers instead of humans, but for the same ultimate purpose of generating new insights that contribute to human knowledge. These technological advances are a boon for the interests served by the freedom to extract. But their status under copyright law is not straightforward. Unlike the human brain, a computer is recognized under the law as a tangible medium in which copies of works can be fixed. This means that a computer that analyzes a corpus of works in order to extract unprotected material typically reproduces those works in copies, an activity that implicates one of the copyright owner's exclusive rights.[134] Because the process involves the copying of works in their

132. *Id.* at 956.

133. *See, e.g.*, DSMC, Inc. v. Convera Corp., 479 F. Supp. 2d 68, 83 (D.D.C. 2007); Madison River Mgmt. Co. v. Bus. Mgmt. Software Corp., 387 F. Supp. 2d 521, 537 (M.D.N.C. 2005).

134. *See generally* Casey & Lemley, *supra* note 3 (acknowledging that machine learning implicates copyright but arguing that it should generally be recognized as fair use); Grimmelmann, *Literate Robots*, *supra* note 47 (describing how courts have deployed fair use to address robotic reading). *But see* Bracha,

entireties for purposes of extracting the unprotected elements, this activity exceeds what the idea/expression distinction and related scope-limiting doctrines permit. Courts have nonetheless recognized the value of extraction even in this new technological context in which extraction requires copying entire works, and they have again deployed fair use to vindicate the freedom to extract.

The Supreme Court has not yet directly addressed the permissibility of text data mining, but it has recognized the general principle that extracting and disseminating *information about* a work should not be the exclusive purview of the copyright owner. So, for example, in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*,[135] the Court suggested that the fair use analysis favors defendants whose purpose is "provid[ing] otherwise unavailable information about the original."[136]

Lower court cases have vindicated this principle in the specific context of text data mining. For example, in *A.V. ex rel. Vanderhye v. iParadigms, LLC*,[137] the Fourth Circuit considered a copyright claim by students whose high school papers were archived in the defendant's plagiarism detection database.[138] Although the defendant copied the plaintiffs' works in their entireties, including both their protected and unprotected elements, the purpose of that copying was to generate *unprotectable information*, specifically, information about whether student papers submitted to the system were original or copied from preexisting works.[139] The court held that this copying was noninfringing fair use.[140] In the course of its analysis, it quoted with approval the district court's observation that "iParadigms' use of the plaintiffs' works 'relate[d] solely to the comparative value of the works.'"[141] That is, iParadigms was copying the works merely to extract from them the facts necessary to operate its plagiarism detection system.

The Second Circuit relied on *iParadigms* and further vindicated the freedom to extract in *Authors Guild v. Google*.[142] There, the court applied the fair use doctrine to hold that Google did not infringe the copyrights in books that it copied for purposes of generating a searchable index of those books.[143] The

Work of Copyright, *supra* note 46 (manuscript at 8) (arguing that generative AI training does not involve reproduction of copyrightable subject matter and so does not amount to prima facie copyright infringement).

135. 598 U.S. 508 (2023).
136. *Id.* at 545.
137. 562 F.3d 630 (4th Cir. 2009).
138. *Id.* at 633.
139. *See id.* at 638–39.
140. *Id.* at 640.
141. *Id.* (alteration in original).
142. 804 F.3d 202 (2d Cir. 2015).
143. *Id.* at 206.

searchable index was a powerful tool for generating *information about* the books—that is, for extracting and disseminating unprotectable facts. That Google necessarily copied more than facts—indeed copied the books in their entireties—did not foreclose the exercise of the freedom to extract.[144] As the court explained:

> Google's making of a digital copy to provide a search function is a transformative use, which augments public knowledge by making available information *about* Plaintiffs' books without providing the public with a substantial substitute for matter protected by the Plaintiffs' copyright interests in the original works or derivatives of them.[145]

When the court focused on Google's copying of entire works, in connection with the third factor of the fair use analysis, it emphasized that the purpose of making those copies was to extract and disseminate unprotected information: "While Google *makes* an unauthorized digital copy of the entire book, it does not reveal that digital copy to the public. The copy is made to enable the search functions to reveal limited, important information about the books."[146]

Authors Guild v. Google built upon previous cases applying fair use to vindicate the practices of search engines. For example, it cited its own related decision in *Authors Guild v. HathiTrust*,[147] in which it had concluded that the nonprofit consortium HathiTrust's copying of books from the collections of member libraries was fair use where the copying was for purposes of creating a full-text search engine (as well as providing access for print-disabled users).[148] The *HathiTrust* district court had come to the same conclusion, and also observed how HathiTrust's search functionality enabled beneficial uses of extracted information beyond mere search—including "new methods of academic inquiry such as text mining."[149]

In *Kelly v. Arriba Soft Corp.*,[150] the Ninth Circuit held that an image search engine that disseminated thumbnail images of the plaintiff's works provided valuable information about those works—information that was not duplicative of the artistic works from which it was extracted:

144. *See* Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 97 (2d Cir. 2014); Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1165 (9th Cir. 2007); Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 610 (2d Cir. 2006); Kelly v. Arriba Soft Corp., 336 F.3d 811, 818 (9th Cir. 2003); Núñez v. Caribbean Int'l News Corp., 235 F.3d 18, 22 (1st Cir. 2000).

145. *Authors Guild*, 804 F.3d at 207.

146. *Id.* at 221–22.

147. 755 F.3d 87, 105 (2d Cir. 2014).

148. *Id.* at 105.

149. Authors Guild, Inc. v. HathiTrust, 902 F. Supp. 2d 445, 460 (S.D.N.Y. 2012), *aff'd in part, vacated in part*, 755 F.3d 87 (2d Cir. 2014).

150. 336 F.3d 811 (9th Cir. 2003).

> [Plaintiff's] images are artistic works intended to inform and to engage the viewer in an aesthetic experience. . . . [Defendant's] use of [Plaintiff's] images in the thumbnails is unrelated to any aesthetic purpose. [Defendant's] search engine functions as a tool to help index and improve access to images on the internet and their related web sites.[151]

The Ninth Circuit considered image search engines' use of thumbnail images again in *Perfect 10, Inc. v. Amazon.com, Inc.*,[152] where it reiterated that an image can be used in a search engine to provide navigational information extracted from its original expressive purpose: "Although an image may have been created originally to serve an entertainment, aesthetic, or informative function, a search engine transforms the image into a pointer directing a user to a source of information."[153]

The Register of Copyrights has also recognized the principle that extracting and disseminating information about a work should not be the exclusive purview of the copyright owner, describing text data mining as an activity that "provide[s] information about works by identifying trends or calculating statistics, which differs from the expressive or informative purposes of the original works."[154] Commentators have likewise concluded that text data mining is generally fair use.[155] This widespread recognition of the legitimacy of text data mining, typically via the application of the fair use doctrine, reflects an appreciation of the value of the freedom to extract—even in cases in which the technology of extraction requires incidental copying of entire copyrighted works.

151. *Id.* at 818.

152. 508 F.3d 1146 (9th Cir. 2007).

153. *Id.* at 1165. An influential district court opinion during this period addressed another feature of search technology. In *Field v. Google Inc*., 412 F. Supp. 2d 1106 (D. Nev. 2006), the court considered Google's use of "cached" snapshots of the web pages it indexes. *Id.* at 1117–18. In holding the dissemination of these cached copies was fair use, the court explained how the copies allow the extraction of several types of information from the copyrighted works. *Id.* at 1117–23.

154. U.S. COPYRIGHT OFF., SECTION 1201 RULEMAKING: EIGHTH TRIENNIAL PROCEEDING TO DETERMINE EXEMPTIONS TO THE PROHIBITION ON CIRCUMVENTION 109 (2021) [hereinafter U.S. COPYRIGHT OFF., SECTION 1201].

155. *See, e.g.*, Sag, *The New Legal Landscape*, *supra* note 3, at 300–01; Carroll, *supra* note 47, at 935; *cf.* ABRAHAM DRASSINOWER, WHAT'S WRONG WITH COPYING? 87 (2015) (arguing that "merely technical reproduction incidental to the operation of digital technology cannot give rise to liability . . . The use is noninfringing because it is noncommunicative. It is not a *fair* use but a *non*use of the work as a work."); Bracha, *Work of Copyright*, *supra* note 46 (manuscript at 26) ("Non-expressive copies involve no enjoyment of any expression qua expression . . . . As a result, the copying does not involve any copyrightable subject matter, and should be found non-infringing long before ever reaching the fair use question.").

d. Photos and Videos as Documentary Evidence of Facts

Several cases about photographs and video footage also grapple with the freedom to extract in the context of copies of entire copyrighted works. In some cases, courts have held that unauthorized copying and distribution of photos or videos is a fair use where the defendant's purpose is to communicate the facts that are embedded in the photo or video. This was the case in *Núñez v. Caribbean International News Corp.*,[156] where a newspaper published photographs of a scantily-clad beauty pageant winner to illustrate stories about the controversy that had erupted about the appropriateness of her posing for the photos.[157] In holding that publication of the photos was fair use, the court emphasized their "informative function" when used to illustrate the news story.[158] Under these circumstances, extracting the newsworthy facts from the copyrighted works required copying the works themselves, because "the pictures were the story," and "[i]t would have been much more difficult to explain the controversy without reproducing the photographs."[159]

In *Elvis Presley Enterprises, Inc. v. Passport Video*,[160] the Ninth Circuit similarly recognized that conveying a factual story can require copying of copyrighted works: "It would be impossible to produce a biography of Elvis without showing some of his most famous television appearances for reference purposes."[161] The court observed that some of the clips were "cited as historical reference points in the life of a remarkable entertainer."[162]

The Ninth Circuit revisited this topic in *Monge v. Maya Magazines, Inc.*,[163] in which the defendant published previously unpublished photos of a clandestine celebrity wedding.[164] The court considered but rejected the argument that publication of the photos was necessary to communicate the facts embedded in them.[165] "Under copyright law, [the defendant] possesses an unfettered right to use any factual information revealed through the photos for the purpose of enlightening its audience, but it can claim no need to bodily appropriate the couple's expression of that information by utilizing portions of

156. 235 F.3d 18 (1st Cir. 2000).

157. *Id.* at 20.

158. *Id.* at 22.

159. *Id.*

160. 349 F.3d 622 (9th Cir. 2003).

161. *Id.* at 629.

162. *Id.* Some of the defendant's use went beyond mere extraction, however. The court explained that "many of the film clips seem to be used in excess of this benign purpose, and instead are simply rebroadcast for entertainment purposes that Plaintiffs rightfully own." *Id.*; *see also* L.A. News Serv. v. Reuters Television Int'l, Ltd., 149 F.3d 987, 993 (9th Cir. 1998), *as amended on denial of reh'g and reh'g en banc* (Aug. 25, 1998).

163. 688 F.3d 1164 (9th Cir. 2012).

164. *Id.* at 1168.

165. *Id.* at 1175.

the actual photos."[166] The court distinguished *Núñez*, in which the photos themselves were the subject of the controversy the defendant newspaper was covering: "In contrast, the controversy here has little to do with photos . . . . The photos were not even necessary to prove [the] controverted fact" of the secret wedding, because "the marriage certificate, which is a matter of public record, may have sufficed to inform the public that the couple kept their marriage a secret for two years."[167]

The *Monge* dissent disagreed, emphasizing how use of the photos enhanced the defendant's ability to disseminate the facts the photos depicted because the couple "were celebrities who carefully concealed their relationship" and "use of the photos was thus integral to exposing to the public the depth of their relationship."[168]

Although they came to different conclusions based on the facts of the case, both the *Monge* majority and dissent acknowledged the possibility that reproduction of copyrighted photos can sometimes be necessary to convey the underlying facts. That is, photos and videos, like the computer software, databases, and literary works described above, can sometimes embed unprotected elements in ways that might threaten the freedom to extract if courts did not carefully deploy limiting doctrines to vindicate that freedom.

e. Works Used as Evidence in Legal Proceedings

A series of cases has demonstrated that copyrighted works may be lawfully reproduced in order to extract facts that are relevant in legal proceedings.[169] The need to present those facts in context can sometimes justify copying works in their entireties.

For example, in *American Institute of Physics v. Winstead PC*,[170] the court held that entire scientific articles could be copied for purposes of establishing the state of the art as relevant to patent proceedings.[171] In the course of its fair use analysis of defendants' reproduction of multiple copies of these "non-patent literature" ("NPL") articles, the court explained that the purpose of such use is "to prove the non-copyrightable ideas, procedures, processes, systems, methods of operation, concepts, principles or discoveries extant in the field."[172] In this context, the literature "is transformed from an item of expressive content to

166. *Id.* (cleaned up).

167. *Id.*

168. *Id.* at 1188 (Smith, J., dissenting) ("Contrary to the majority's contentions, a mere marriage certificate would not suffice.").

169. *See* Pamela Samuelson, *Unbundling Fair Uses*, 77 FORDHAM L. REV. 2537, 2592–97 (2009) (discussing cases allowing uses for litigation and other government purposes).

170. No. 3:12-CV-1230-M, 2013 WL 6242843 (N.D. Tex. Dec. 3, 2013).

171. *Id.* at *13; *see also* D.R. Jones, *Law Firm Copying and Fair Use: An Examination of Different Purpose and Fair Use Markets*, 56 S. TEX. L. REV. 313, 315–24 (2014).

172. *Am. Inst. of Physics*, 2013 WL 6242843, at *5.

evidence of the facts within it; the expressive content becomes merely incidental."[173] The court therefore held that the defendants' use of the articles was fair "despite the fact that they are exact copies."[174]

Faced with very similar facts (in a case brought by the same publisher), the court in *American Institute of Physics v. Schwegman, Lundberg & Woessner, P.A.*[175] also held that the defendant law firm's copying of scientific articles was fair.[176] It concluded that the first factor favored fair use because the defendant's purpose was different from the purpose for which the articles were originally prepared.[177] Specifically, the defendant's "allegedly infringing internal copying of the Articles took on an *evidentiary* character."[178]

The *Schwegman* court went on to explain how that evidentiary character of the defendant's use related to the extraction of unprotectable elements from the underlying work: "The Articles are useful to [the defendant] and to the various patent offices as comparative references for the specific inventions that are the subject of pending patent applications, and the *facts* and *ideas* reflected in the Articles are of use to [the defendant], not the Articles' copyrightable manner of expression."[179] As the court noted, this use did not easily fit within the category of "transformative use" that the Supreme Court has repeatedly favored it its fair use analysis.[180] And it is not exactly "non-expressive" either,[181] insofar as the defendant's use communicated to readers the expression in the verbatim copies it made. What was critical to the court, and consistent with my thesis, was that the defendant's purpose was to *extract* facts and ideas (including the fact that the articles formed part of the body of scientific literature that constituted the relevant prior art).

173. *Id.* at *4.

174. *Id.* at *6. The court applied this reasoning not only to the copies submitted to the USPTO, but also to copies "made for attorneys to allow Defendants to determine whether a particular NPL article must be submitted to the USPTO and copies forwarded to clients as attachments to USPTO filings and correspondence." *Id.*

175. No. CIV. 12-528 RHK/JJK, 2013 WL 4666330 (D. Minn. Aug. 30, 2013).

176. *Id.* at *2.

177. *Id.* at *10.

178. *Id.* at *12 (emphasis added).

179. *Id.*

180. The court observed that the

> lack of alteration [of the articles] may make the label "transformative use" a messy fit . . . since the "transformative use" label is most apt when a secondary work created by an alleged infringer actually alters the content of a copyrighted work or incorporates that content into a new work, such as a parody.

Id. at *11.

181. *Cf. supra* notes 47–49 and accompanying text (discussing the concept of "non-expressive" use).

Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey[182] provides another example of extraction of facts for evidentiary purposes.[183] The court rejected a claim alleging that a law firm had infringed the plaintiff's copyright by reproducing and displaying archived copies of the plaintiff's website (in the course of trademark and copyright litigation that the plaintiff had instigated against a client of the firm).[184] The plaintiff alleged the defendant infringed by displaying the website images on firm computers and printing out copies of the images after accessing the images using the Internet Archive's "Wayback Machine."[185] The court rejected those claims on the basis of fair use, explaining the purpose of the defendant's use this way:

> The [defendant's] purpose in viewing and printing copies of the archived images of [the plaintiff's] website was primarily to defend their clients. The [defendant] viewed these archived web pages to assess the merit of the claims brought against their client. They hoped they might discover facts allowing them to refute the allegations.[186]

In other words, the defendant needed to view and copy the work in order to extract facts from the work and also to document the existence of those facts. The court held that this purpose favored fair use, as "[i]t would be an absurd result if an attorney defending a client against charges of trademark and copyright infringement was not allowed to view and copy publicly available material, especially material that his client was alleged to have infringed."[187]

Finally, consider *Bond v. Blum*,[188] in which the court allowed copying of an entire incriminating autobiographical manuscript (in which the plaintiff described evading punishment for remorselessly murdering his father) for purposes of establishing the author's lack of fitness in a child custody proceeding.[189] In discussing the defendants' purpose in copying the manuscript, the court emphasized the need to extract its facts for their evidentiary value:

> [T]he narrow purpose of defendants' use of the manuscript is for the evidentiary value of its content insofar as it contains admissions that [the plaintiff] may have made against his interest when he bragged about his conduct in murdering his father, in taking advantage of the juvenile justice system, and in benefiting from his father's estate. These are all

182. 497 F. Supp. 2d 627 (E.D. Pa. 2007).

183. *See id.* at 631.

184. *Id.* at 629, 631–32.

185. *Id.* at 635. The plaintiff also alleged that the defendant had saved copies on firm hard drives and distributed copies to another law firm, but the court held that the plaintiff presented no evidence supporting these allegations. *Id.* at 639.

186. *Id.* at 636.

187. *Id.* at 637.

188. 317 F.3d 385 (4th Cir. 2003).

189. *Id.* at 389.

> facts relevant to the custody decision, and their use does not draw on [the plaintiff's] mode of expression.[190]

Under these circumstances, the need to use the manuscript for its factual, evidentiary value justified copying it in its entirety—including even its expressive elements.[191] "The use of the copyrighted material in this context, even the entire manuscript, does not undermine the protections granted by the Act but only serves the important societal interest in having evidence before the factfinder."[192]

Notice how these cases add an additional layer to the notion of extraction. The courts looked favorably on the uses not only because their purpose was to extract facts from the copyrighted works, but also because the very existence of the works constituted a valuable fact that helped to justify copying each work as a whole. In the patent prosecution cases, it was important to know that the articles at issue existed as part of the scientific literature. In *Healthcare Advocates*, it was important to know how the website at issue appeared during a particular time period in order to assess trademark and copyright claims. In *Bond*, it was important in the child custody case to know not only the facts depicted in the manuscript (some of which were matters of public record), but the fact *of* the manuscript—that its author bragged in writing about the horrific facts it described. This is yet another scenario in which it may be necessary to use an entire work in order to effectively extract something unprotected from the work. In this case, the work in its entirety is necessary to accurately communicate the existence and character of the work itself, which can operate as a fact for purposes of a legal proceeding.

5. The Useful Article Doctrine

The Copyright Act's useful article doctrine facilitates extraction of unprotectable elements in the specific context of "useful articles," which the statute defines as "article[s] having an intrinsic utilitarian function that is not merely to portray the appearance of the article or to convey information."[193] The statute provides that "the design of a useful article" shall be protected as a "pictorial, graphic, or sculptural work only if, and only to the extent that, such design incorporates pictorial, graphic, or sculptural features that can be

190. *Id.* at 395.

191. *Id.* at 396 ("It is conceded that the defendants' challenged use of the manuscript in the state-court proceeding involved all, or nearly all, of the copyrighted work. Its use, however, was not for its expressive content, but rather for its allegedly factual content. The sole purpose and intent of introducing Bond's manuscript was to obtain admissions of fact against his interest in an effort to prove that his home would not be a suitable place for custody of children.").

192. *Id.*

193. 17 U.S.C. § 101. The definition goes on to provide that "[a]n article that is normally a part of a useful article is considered a 'useful article.'" *Id.*

identified separately from, and are capable of existing independently of, the utilitarian aspects of the article."[194] So, for example, the shape of a shoe can be protected only if—and only to the extent that—features of the shape can be identified that are separable from those features that make the shoe perform as a shoe.

This special provision for useful articles can be understood as offering an additional layer of protection for the freedom to extract. Even without the doctrine, works with intrinsic utilitarian functions should receive thin protection that does not extend to their functional characteristics—which amount to unprotectable methods of operation. However, as some of the illustrations above demonstrate more generally, it can be hard to distinguish and surgically extract only unprotectable elements. The useful article doctrine seems to anticipate this difficulty and nip it in the bud by denying even thin copyright protection to useful articles unless the protected elements can be identified separately. In other words, even seemingly protectable aesthetic elements of a useful article should not be protected if those elements would be too difficult to extract.

In practice, application of the concept of "separability" has proven as difficult as any act of distinguishing protected from unprotected elements, thus limiting the doctrine's utility as a bulwark of the freedom to extract.[195] There are examples, however, where the doctrine has successfully served this purpose. Consider *Dennis v. Nike, Inc.*[196] There, the plaintiff claimed that Nike had infringed his design for an athletic shoe with springs in the sole arranged in a square formation.[197] The court dismissed the plaintiff's complaint for several overlapping reasons.[198] In its analysis of the useful article doctrine, the court made clear that this doctrine would have been an independent basis for dismissing the complaint even if the allegedly infringed design had not been filtered out by other scope-limiting doctrines.[199] In other words, the useful

194. *Id.*

195. Critiques of the Supreme Court's most recent opinion applying the doctrine, *Star Athletica, L.L.C. v. Varsity Brands, Inc.*, 580 U.S. 405 (2017), have been pointed and widespread. *See, e.g.*, Rebecca Tushnet, *Shoveling a Path After* Star Athletica, 66 UCLA L. REV. 1216, 1218, 1221 (2019); Jane C. Ginsburg, *The Sum Is More Public Domain than Its Parts?: U.S. Copyright Protection for Works of Applied Art Under* Star Athletica*'s Imagination Test*, 166 U. PA. L. REV. ONLINE 83, 91–100 (2017); Christopher Buccafusco & Jeanne C. Fromer, *Forgetting Functionality*, 166 U. PA. L. REV. ONLINE 119, 122 (2017); Barton Beebe, Star Athletica *and the Problem of Panaestheticism*, 9 U.C. IRVINE L. REV. 275, 278–79 (2019); Mala Chatterjee, *Conceptual Separability as Conceivability: A Philosophical Analysis of the Useful Articles Doctrine*, 93 N.Y.U. L. REV. 558, 584–86 (2018); Tyler T. Ochoa, *What Is a "Useful Article" in Copyright Law After* Star Athletica*?*, 166 U. PA. L. REV. ONLINE 105, 115, 117 (2017); Mark P. McKenna, *Knowing Separability When We See It*, 166 U. PA. L. REV. ONLINE 127, 128–30 (2017).

196. No. 2:22-CV-04515-SB-PD, 2023 WL 2356719 (C.D. Cal. Feb. 13, 2023).

197. *Id.* at *4.

198. *See id.* at *3.

199. *See id.* at *4–5.

article doctrine operated in this case as a backstop that ensured the freedom to extract unprotectable utilitarian elements even if they were too entangled with expressive elements to be surgically extracted. As the court explained:

> Even if the Court were to accept Plaintiff's contention that the scope of his copyrighted work includes a design with springs in square formation, Plaintiff admits that they serve a utilitarian purpose, "impact absorption." . . . Plaintiff has not identified any element of his design that can be separated from the utilitarian purpose of the springs in the sole of the shoe so as to be protectable in copyright.[200]

Beyond serving as a backstop that ensures the freedom to extract even (indeed, especially) when it is difficult to separate protectable and unprotectable elements, the useful article doctrine puts the burden of proof on the plaintiff (to prove separability) in circumstances in which the threat to the freedom to extract unprotectable functional elements might be especially acute.[201]

6. Functionality of Computer Programs

Another provision of the Copyright Act ensures that owners of copies of computer software can extract the functionality from that software by running it. 17 U.S.C. § 117 specifies, inter alia:

> Notwithstanding the provisions of section 106, it is not an infringement for the owner of a copy of a computer program to make or authorize the making of another copy or adaptation of that computer program provided . . . that such a new copy or adaptation is created as an essential step in the utilization of the computer program in conjunction with a machine and that it is used in no other manner.[202]

This special explicit protection for the freedom to extract is necessary due to the nature of computer programs, which typically need to be copied into temporary computer memory in order to run. In the absence of section 117, owners of copies of computer programs might be unable lawfully to use those programs for the functional purposes for which they were purchased without permission from the copyright owner.

200. *Id.* at *4; *see also* Ross v. Apple, Inc., 741 F. App'x 733, 737 (11th Cir. 2018); *cf.* Oren Bracha & John M. Golden, *Redundancy and Anti-Redundancy in Copyright*, 51 CONN. L. REV. 247, 256 (2019) ("[C]opyright law features multiple sets of doctrines that address the same or overlapping concerns in various circumstances or otherwise perform overlapping functions.").

201. *See* Jane C. Ginsburg, *"Courts Have Twisted Themselves into Knots": U.S. Copyright Protection for Applied Art*, 40 COLUM. J.L. & ARTS 1, 18 n.71 (2016) ("The burden of proving separability falls on the plaintiff; many cases have been dismissed at the outset for failure to plead with particularity which elements of the article were conceptually separable and why.").

202. 17 U.S.C. § 117(a)(1).

The National Commission on New Technological Uses of Copyrighted Works ("CONTU")[203] report recommending adoption of the current version of section 117 explained that "the law should provide that persons in rightful possession of copies of programs be able to use them freely without fear of exposure to copyright liability."[204]

By adopting section 117 and thereby ensuring that owners of copies of computer programs can in fact *use* them, Congress aimed to protect the freedom to extract unprotectable functionality, which might otherwise be subject to copyright owners' veto based on the technological happenstance that using a computer program requires copying it (that is, copying it in its entirety, including both functional and expressive elements). This protection of the freedom to extract has been weakened in some cases, however, by enforcement of licenses purporting to deny software users the status of copy owners and thus to deny them the rights section 117 grants to "the owner of a copy."[205]

7. Making Unprotected Elements Extractable for People with Print-Related Disabilities

17 U.S.C. § 121 (known as the Chafee Amendment) permits a statutorily designated "authorized entity"[206] to reproduce copyrighted works for purposes of making those works accessible to people who are blind or have other print-related disabilities.[207] Without this provision, copyright law could be deployed to deny people with disabilities access to both the protectable and unprotectable elements of copyrighted works. This provision can therefore be understood as another way in which current law safeguards the freedom to extract—although it does more than this (by promoting access to entire works) and less (because it does not ensure that such access is actually accomplished).[208]

203. Congress created the National Commission on New Technological Uses of Copyrighted Works ("CONTU") in 1974 to advise it on how to address new technological developments relevant to copyright. Act of Dec. 31, 1974, Pub. L. No. 93-573, sec. 201, 88 Stat. 1873, 1873–74.

204. NATIONAL COMMISSION ON NEW TECHNOLOGICAL USES OF COPYRIGHTED WORKS (CONTU), FINAL REPORT 13 (1978).

205. *See, e.g.*, MAI Sys. Corp. v. Peak Comput., Inc., 991 F.2d 511, 518 n.5 (9th Cir. 1993) (holding that section 117 did not apply where copyright owner had "licensed" its software to customers who therefore did not qualify as "owners"). For critique of *MAI* and its progeny, see, for example, Brian W. Carver, *Why License Agreements Do Not Control Copy Ownership: First Sales and Essential Copies*, 25 BERKELEY TECH. L.J. 1887, 1899–900 (2010); 2 MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 8.08.

206. The statute defines "authorized entity" to mean "a nonprofit organization or a governmental agency that has a primary mission to provide specialized services relating to training, education, or adaptive reading or information access needs of blind or other persons with disabilities." 17 U.S.C. § 121(d)(2).

207. *Id.*

208. *See* Blake E. Reid, *Copyright and Disability*, 109 CALIF. L. REV. 2173, 2198–200 (2021) (describing the history, purpose, provisions, and shortcomings of the Chafee Amendment); Blake E.

When introducing the original version of 17 U.S.C. § 121, Senator Chafee explained its purpose to ensure timely access to information, ideas, and facts—that is, to the unprotectable elements contained in inaccessible protected works. Senator Chafee referred, in particular, to problems encountered in the production of braille textbooks for blind students, explaining that "[t]he amendment . . . seeks to end the unintended censorship of blind students' access to current *information*. . . . Passage of this amendment will permit the speedy access to *information* that blind people need."[209]

Like computer object code that can obscure unprotectable elements in a way that can justify copying necessary to extract those elements, inaccessible formats can obscure unprotectable elements from those with relevant disabilities, justifying copying necessary to communicate unprotectable facts and ideas. Section 121 is thus another example of how the Copyright Act safeguards the freedom to extract.[210]

8. Beyond Copyright Infringement: Extraction Through Circumvention of Technological Protection Measures

Section 1201 of the Copyright Act was enacted in 1998 as part of the Digital Millennium Copyright Act ("DMCA").[211] Its purpose was to provide a legal backstop to technological protection measures ("TPMs") deployed by and on behalf of copyright owners to inhibit infringement of copyrighted works. Such measures can sequester both protected and unprotected elements (and even works in the public domain), and thus have the potential to hinder the freedom to extract.[212] But several courts have interpreted the DMCA to avoid that result.

TPMs include access control technology—tools that determine who can have access to the works to which the technology is applied. Password systems

Reid, *What Copyright Can't Do*, PEPP. L. REV. (forthcoming 2025), https://ssrn.com/abstract=4766540 [https://perma.cc/WH96-AXZF (staff-uploaded archive)] (describing structural limits on copyright law's ability to ensure accessibility of creative works).

209. 142 CONG. REC. 19674 (1996) (emphasis added).

210. Congress and courts have explained that fair use can serve the same purpose. *See, e.g.*, H.R. REP. NO. 94-1476, at 73 (1976) ("[T]he making of a single copy or phonorecord by an individual as a free service for a blind persons [sic] would properly be considered a fair use under section 107."); Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 455 n.40 (1984) ("Making a copy of a copyrighted work for the convenience of a blind person is expressly identified by the House Committee Report as an example of fair use, with no suggestion that anything more than a purpose to entertain or to inform need motivate the copying."); Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 103 (2d Cir. 2014) ("Weighing the [fair use] factors together, we conclude that the doctrine of fair use allows the Libraries to provide full digital access to copyrighted works to their print-disabled patrons.").

211. Digital Millennium Copyright Act, Pub. L. No. 105-304, 112 Stat. 2860 (1998) (codified as amended in scattered sections of 17 U.S.C.).

212. *See* Reichman et al., *supra* note 88, at 1019 (expressing concern, in light of the DMCA, "with gaining access to copyrighted works in the digital environment in order to extract unprotectable subject matter, such as ideas and disparate facts").

used to control access to music platforms are a commonplace example of access control technology. Another type of TPM controls the actions that are possible once a user has access to a work. Technology that prevents even authorized users from printing copies of e-books is an example of this type of TPM. Either type of TPM can make it difficult to exercise the freedom to extract. For example, access control technology designed to prevent a video game console from being used with games made by an unauthorized manufacture (by restricting access to the console by software that does not include a secret lock-out code) can prevent extraction of the type of interface specification subject to the freedom to extract in cases like *Sega* and *Sony*. Or technology that prevents copying can make it impossible to duplicate software for the archival purpose expressly permitted by 17 U.S.C. § 117.[213]

The DMCA added several TPM-related provisions to the Copyright Act. The provision codified at 17 U.S.C. § 1201 is entitled "Circumvention of copyright protection systems."[214] It addresses circumvention of both technological measures that control access to works protected by the Copyright Act and technological measures that protect a right of a copyright owner.[215] As to the former, the statute prohibits circumvention of such measures and trafficking in technology designed to circumvent such measures.[216] As to the latter, the statute prohibits only trafficking in circumvention technology.[217]

A trio of cases decided in the first decade following the DMCA's enactment grappled with its potential impact on the freedom to extract.

First, in *Chamberlain Group, Inc. v. Skylink Technologies, Inc.*,[218] the Federal Circuit resolved a dispute between a manufacturer of automatic garage door openers and a manufacturer of universal remote control devices designed to work with those garage door openers.[219] The plaintiff alleged that the defendant had violated the DMCA by trafficking in technology designed to circumvent the access control embedded in the plaintiff's garage door openers.[220]

The Federal Circuit rejected that argument, first emphasizing that the DMCA should be interpreted to prohibit "only forms of access that bear a reasonable relationship to the protections that the Copyright Act otherwise affords copyright owners."[221] That interpretation, requiring a "nexus between access and protection,"[222] doomed the plaintiff's argument, because the

213. *See* Vault Corp. v. Quaid Software Ltd., 847 F.2d 255, 262 (5th Cir. 1988).
214. Digital Millennium Copyright Act § 1201.
215. 17 U.S.C. § 1201.
216. *Id.* § 1201(a)(1)(A), (a)(2).
217. *Id.* § 1201(b)(1).
218. 381 F.3d 1178 (Fed. Cir. 2004).
219. *Id.* at 1178.
220. *Id.* at 1183.
221. *Id.* at 1202.
222. *Id.* at 1204.

Copyright Act does not prohibit lawful owners of garage door openers from accessing the software embedded in those openers and "Chamberlain neither alleged copyright infringement *nor explained how the access provided by the [defendant's] transmitter facilitates the infringement of any right that the Copyright Act protects*."[223] Skylink could therefore not be held liable for trafficking in devices that merely facilitated that access.[224]

The *Chamberlain* court's refusal to interpret the DMCA to forbid trafficking in circumvention tools that merely facilitate lawful access to copyrighted works is generally consistent with cases like *Sega*, *Sony*, and *WIREData*, holding that copyright owners cannot prohibit access to unprotected elements of copyrighted works even if those elements are locked within expressive elements that need to be copied in order to exercise the freedom to extract. That said, the court did not address arguments that would have been even more squarely in line with the freedom to extract. In particular, the defendant in *Chamberlain* had argued that Chamberlain's access control technology should not receive DMCA protection because it "does not protect a copyrighted computer program, but instead protects an uncopyrightable process."[225] In other words, the defendant argued that its freedom to extract uncopyrightable elements—here, the interface that would allow competing devices to open owners' garage doors—should not be obstructed by a technological lock. The court also found no need to decide the defendant's argument that its activities were shielded by 17 U.S.C. § 1201(f)'s exception for reverse engineering to achieve interoperability.[226]

The Sixth Circuit considered a similar scenario in *Lexmark International, Inc. v. Static Control Components, Inc.* As described above, in *Lexmark* the Sixth Circuit considered printer manufacturer Lexmark's claims that Static Control Components infringed the copyright in Lexmark's toner loading program when Static Control manufactured chips that copied that program for the purpose of allowing remanufactured toner cartridges to function in Lexmark's printers.[227] In addition to the copyright infringement claim, Lexmark alleged that Static Control violated the DMCA by trafficking in technology that circumvented Lexmark's access control measures.[228] Lexmark's theory was that the Static Control chips circumvented Lexmark's authentication sequence, and thereby provided unauthorized access to both the Toner Loading Program embedded in Lexmark cartridges and the Printer Engine Program embedded in Lexmark

223. *Id.*

224. *Id.*

225. *Id.* at 1186.

226. *Id.* at 1200 n.15.

227. Lexmark Int'l, Inc. v. Static Control Components, Inc., 387 F.3d 522, 522 (6th Cir. 2004), *abrogated on other grounds by* eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388 (2006).

228. *Id.* at 531.

printers.[229] Echoing *Chamberlain*, the Sixth Circuit held that the DMCA is not violated by technology that merely provides consumers with access to works to which they already have lawful access.[230] The court arrived at this conclusion by determining that Lexmark's authentication sequence was not in fact a technological protection measure safeguarded by the DMCA because it did not "effectively control[] access to" a copyright-protected work as specified by the DMCA.[231] For example, with regard to the Printer Engine Program, the court observed that "[i]t is not Lexmark's authentication sequence that 'controls access' to the Printer Engine Program. It is the purchase of a Lexmark printer that allows 'access' to the program."[232]

In light of this logic (which the court extended to the Toner Loading Program), the Sixth Circuit did not squarely address Static Control's section 1201(f) argument, although it did express disagreement with the district court's reasons for disallowing the defense.[233] Neither did the court squarely address the argument that the elements protected by the (purported) technological protection measure were unprotectable subject matter. Its holding nonetheless had the practical effect of vindicating the freedom to extract those elements, as did the Federal Circuit's subsequent decision in *Storage Technology Corp. v. Custom Hardware Engineering & Consulting, Inc.*,[234] in which it followed its previous holding in *Chamberlain* to hold that the DMCA requires a nexus between the alleged circumvention and copyright infringement.[235]

These cases demonstrate how technological protection measures (and the DMCA provisions designed to reinforce them) have the potential to endanger the freedom to extract. Some courts have avoided this danger by recognizing that the law should not prohibit circumvention that facilitates behavior that copyright law does not otherwise forbid—be it lawful consumer use of equipment they have purchased or lawful third-party repair or interoperability with that equipment. This reasoning could similarly apply where circumvention aims to extract unprotected elements from protected works. Not all courts have interpreted the DMCA's prohibitions in a way that safeguards the freedom to extract, however. In *MDY Industries, LLC v. Blizzard Entertainment, Inc.*,[236] for

229. *Id.*

230. *Id.* at 564.

231. *Id.* at 528 (citing 17 U.S.C. § 1201).

232. *Id.* at 546 (citation omitted); *see also id.* at 546–47 ("Anyone who buys a Lexmark printer may read the literal code of the Printer Engine Program directly from the printer memory, with or without the benefit of the authentication sequence, and the data from the program may be translated into readable source code after which copies may be freely distributed. No security device, in other words, protects access to the Printer Engine Program Code and no security device accordingly must be circumvented to obtain access to that program code." (citation omitted)).

233. *Id.* at 550–51.

234. 421 F.3d 1307 (Fed. Cir. 2005).

235. *Id.* at 1319.

236. 629 F.3d 928 (9th Cir. 2010).

example, the Ninth Circuit rejected *Chamberlain*'s nexus requirement as inconsistent with the text of the statute,[237] observing that "[i]f greater protection of the public's ability to access copyrighted works is required, Congress can provide such protection by amending the statute."[238]

Even where courts do not recognize a nexus requirement for violations of the DMCA's provisions, there is another DMCA safety valve that recognizes, in a piecemeal way, the importance of safeguarding the freedom to extract against overreaching technological protection measures. In addition to the recognition of reverse engineering for interoperability in 17 U.S.C. § 1201(f), mentioned above,[239] section 1201(a)(1)(C) of the DMCA authorizes the Librarian of Congress (upon recommendation of the Register of Copyrights) to grant renewable three-year exceptions to the prohibition of circumvention of access controls to certain users. Specifically, these exemptions may be granted to users for whom the prohibition would adversely affect their ability to make noninfringing use of a particular class of works.[240]

In implementing this provision, the Library of Congress has recognized the importance of certain types of extraction, granting an exception for extraction in the form of text data mining that might otherwise be obstructed by access control technology. The exception applies to circumvention of access control measures on motion pictures and electronically-distributed literary works that are accessed by researchers at nonprofit higher education institutions "solely to deploy text and data mining techniques on a corpus" of such works "for the purpose of scholarly research and teaching."[241] In recommending the adoption of these exemptions,[242] the Register of Copyrights explained that text data mining "involves digitizing and downloading large numbers of works to create datasets on which researchers can perform algorithmic extractions to investigate questions and observe trends."[243] She explained that the research enabled by the exemption would "provide information about works by identifying trends or calculating statistics, which differs from the expressive or informative purposes of the original works."[244]

237. *Id.* at 948–52.

238. *Id.* at 951.

239. Additional exemptions include, for example, 17 U.S.C. § 1201(j) (security testing).

240. *Id.* § 1201(a)(1)(C); *see also* Green v. U.S. Dep't of Just., 111 F.4th 81, 101 (D.C. Cir. 2024) ("It is generally fair to say that what the fair use defense does for copyright infringement, the exemptions do for section 1201(a).").

241. 37 C.F.R. § 201.40(b)(4)–(5).

242. U.S. COPYRIGHT OFF., SECTION 1201 RULEMAKING, *supra* note 154, at 111.

243. *Id.* at 102.

244. *Id.* at 109. These exemptions were recently renewed as part of the ninth triennial proceeding. *See* Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies, 89 Fed. Reg. 85437-01 (2024).

Because the DMCA regulates the circumvention of tools that can control access to and use of unprotected elements of copyrighted works, it has the potential to limit the freedom to extract. Some courts and the Copyright Office have resisted interpreting the DMCA so broadly, however, thus vindicating the freedom to extract.

9. Other Examples Beyond Copyright

The protections and exceptions of the Copyright Act interact with several other bodies of law, including contract law, computer fraud law, and trademark law. These adjacent regimes sometimes reinforce—and sometimes interfere with—the freedom to extract.

One example is contract law. Copyright owners frequently try to augment the rights granted under copyright, and undercut the limitations of copyright, by imposing additional contractual restrictions on how their works may be used. Several cases have tested the limits on parties' ability to forbid by contract the exercise of the freedom to extract under copyright law.

The Fifth Circuit encountered this issue in *Vault Corp. v. Quaid Software Ltd.*[245] Plaintiff Vault produced "PROLOK" computer diskettes designed to prevent unauthorized copying of the computer software stored on them.[246] Vault sold these diskettes to software companies that wanted to prevent their customers from buying one copy of their programs and then duplicating that copy for more widespread use.[247] Defendant Quaid produced a product, called RAMKEY, that circumvented the PROLOK copy-control system, "facilitat[ing] the creation of a fully functional copy of a program placed on a PROLOK diskette."[248]

Quaid had copied the PROLOK software onto a computer in the process of making the RAMKEY product.[249] The Fifth Circuit held that this copying, "for the express purpose of devising a means of defeating [PROLOK's] protective function,"[250] was not infringing. This copying was within the exception provided by 17 U.S.C. § 117 for copying as an "essential step" in the utilization of a computer program.[251]

245. 847 F.2d 255 (5th Cir. 1988).

246. *Id.* at 256.

247. As the court explained, "[t]he fact that a fully functional copy of a program cannot be made from a PROLOK diskette prevents purchasers from buying a single program and making unauthorized copies for distribution to others." *Id.* at 256–57.

248. *Id.* at 257.

249. *Id.*

250. *Id.* at 261.

251. *Id.* The court also held that Quaid did not directly or indirectly infringe by producing and distributing RAMKEY–which incorporated only a tiny percentage of the PROLOK code (too little to render it substantially similar and thus infringing) and was capable of substantial noninfringing uses (such as making lawful archival copies). *Id.* at 267–68.

The court then went on to consider Vault's claim that Quaid had violated state law by breaching the terms of the license agreement under which PROLOK diskettes were distributed, which provided that "[y]ou may not . . . copy, modify, translate, convert to another programming language, decompile or disassemble" the software embedded on those diskettes.[252] The Louisiana Software License Enforcement Act[253] explicitly provided for enforcement of software license terms that prohibited "reverse engineering, decompilation or disassembly."[254] The Fifth Circuit nonetheless refused to enforce these license terms, explaining that the Louisiana Act was preempted by the federal Copyright Act, insofar as the state law permitted prohibitions on copying that would be permissible under 17 U.S.C. § 117.

Although the *Vault v. Quaid* court did not explain its reasoning in a way that expressly connects to the freedom to extract, the holding and its underlying logic clearly support that freedom. Without preemption or some other limiting doctrine, section 117 and other aspects of copyright law that promote the freedom to extract could be undermined by the enforcement under state law of license provisions that forbid what the Copyright Act permits.[255] Reading *Vault* in light of later cases like *Sega* makes it clear how the preemption of state laws (and state law enforcement of private licenses) that forbid copying even for purposes of reverse engineering might be necessary to ensure that copyright owners cannot lock up unprotected functional elements.[256]

More recently, the district court in *X Corp. v. Bright Data Ltd.*[257] recognized the potential tension between state contract law and copyright law's exclusion of unprotected elements.[258] The case involved a number of state law causes of action by plaintiff X (formerly Twitter) based on defendant Bright Data's scraping and selling of publicly available data from X's social media platform in

252. *Id.* at 269 (alteration in original).

253. Act of July 13, 1984, ch. 27, 1984 La. Acts 1846 (codified at LA. STAT. ANN. § 51:1964).

254. *Vault*, 847 F.2d at 268–69 (citing LA. STAT. ANN. § 51:1964 (2024)).

255. This general line of reasoning was embraced by the district court in *C.B.C. Distribution & Marketing, Inc. v. Major League Baseball Advanced Media, L.P.*, 443 F. Supp. 2d 1077 (E.D. Mo. 2006), *aff'd on other grounds*, 505 F.3d 818 (8th Cir. 2007), in which the court refused to enforce an agreement governing use of baseball players' names and records for purposes of a fantasy baseball game, explaining that "'the strong federal policy favoring the full and free use of ideas in the public domain' as manifested in the laws of intellectual property prevails over the challenged contractual provisions." *Id.* at 1106–07 (quoting Lear, Inc. v. Adkins, 395 U.S. 653, 674 (1968)).

256. *See* Samuelson & Scotchmer, *supra* note 94, at 1630 ("To the extent that enforcement of anti-reverse-engineering clauses would have a detrimental effect on competitive development and innovation, legal decisionmakers may be justified in not enforcing them"); Noam Shemtov, *Circumventing the Idea/Expression Dichotomy: The Use of Copyright, Technology and Contract to Deny Access to Ideas*, *in* EMERGING ISSUES IN INTELLECTUAL PROPERTY 88, 105–08 (Guido Westkamp ed., 2007) (critiquing enforcement of anti-decompilation contractual provisions).

257. No. C 23-03698, 2024 WL 2113859 (N.D. Cal. May 9, 2024).

258. *Id.* at *12.

violation of X's terms of service.[259] Noting in connection with the breach of contract claim that this was not a case of "an arm's length contract between two sophisticated parties in which one or the other adjusts their rights and privileges under federal copyright law," but rather "a massive regime of adhesive terms imposed by X Corp.," the court held that the claims based on Bright Data's scraping and selling of data were preempted by the Copyright Act.[260] X Corp's theory of state law liability, the court reasoned, would interfere with the Copyright Act's placement of unprotected elements in the public domain, giving itself "de facto copyright ownership over content that Congress declined to extend copyright protection to in the first place."[261] The court rejected this attempt to prohibit extraction of unprotected elements, holding X's state law claims based on scraping preempted because they would undermine federal law.[262] Thus, as in *Vault*, the court used preemption to vindicate the freedom to extract.

Not all courts have reinforced the freedom to extract in this way, however. For example, in *Bowers v. Baystate Technologies, Inc.*,[263] the Federal Circuit held that a shrink-wrap license forbidding reverse engineering was not preempted by the Copyright Act, despite acknowledging the continued validity of its previous holding in *Atari Games Corp. v. Nintendo of America, Inc.* that "reverse engineering object code to discern the unprotectable ideas in a computer program is a fair use."[264] The court distinguished *Vault v. Quaid* (inaccurately[265]) as involving preemption of "a state law prohibiting all copying of a computer program" as opposed to "private contractual agreements supported by mutual assent and consideration."[266]

Judge Dyk dissented on the question of preemption, recognizing the square conflict between the majority's opinion and *Vault v. Quaid*[267] and the

259. *Id.* at *3.

260. *Id.* at *12.

261. *Id.* at *14 ("X Corp. would upend the careful balance Congress struck between what copyright owners own and do not own, and what they leave for others to draw on. . . . This shrinks the public domain, restricting free reproduction, adaptation, distribution, and display of publicly available, non-expressive material.").

262. *Id.* at *13–14.

263. 320 F.3d 1317 (Fed. Cir. 2003).

264. *Id.* at 1325 (citing Atari Games Corp. v. Nintendo of Am., Inc., 975 F.2d 832, 843 (Fed. Cir. 1992)).

265. *See id.* at 1337 (Dyk, J., concurring in part and dissenting in part) ("From a preemption standpoint, there is no distinction between a state law that explicitly validates a contract that restricts reverse engineering ([as in Vault Corp. v. Quaid Software, Ltd., 847 F.2d 255 (5th Cir. 1988)]) and general common law that permits such a restriction (as here).").

266. *Id.* at 1325 (majority opinion).

267. *Id.* at 1335 (Dyk, J., concurring in part and dissenting in part) ("By holding that shrinkwrap licenses that override the fair use defense are not preempted by the Copyright Act, the majority has rendered a decision in conflict with the only other federal court of appeals decision that has addressed the issue—the Fifth Circuit decision in *Vault Corp. v. Quaid Software Ltd.*" (citations omitted)).

threat to copyright policy posed by the majority's approach, which "permits state law to eviscerate an important federal copyright policy reflected in the fair use defense" and "threatens other federal copyright policies as well."[268]

Judge Dyk articulated the copyright policies at issue in terms of the freedom to extract unprotected elements even when they are embedded in unreadable computer code. He explained that without a fair use exception for reverse engineering, copyright owners might be able to "achieve protection for an idea simply by embodying it in a computer program."[269] Judge Dyk would have recognized contractual override of this freedom to extract via reverse engineering only in cases of "freely negotiated" contracts, but not for shrinkwrap contracts of adhesion.[270]

Another source of law that interacts with the freedom to extract is the Computer Fraud and Abuse Act ("CFAA").[271] This federal law authorizes civil and criminal penalties for anyone who "intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains . . . information from any protected computer,"[272] with "protected computer" defined to include any computer "used in or affecting interstate or foreign commerce or communication"[273] (and thus including any computer connected to the internet). If interpreted broadly, this prohibition could be applied to restrict the freedom to extract. Indeed, operators of internet platforms have argued, sometimes successfully, that the CFAA prohibits unauthorized collection of data from a publicly available website if that access violates the website's terms of service or continues in the face of a cease-and-desist letter,

268. *Id.*

269. *Id.* at 1336 ("[A]n author cannot achieve protection for an idea simply by embodying it in a computer program. . . . Thus, the fair use defense for reverse engineering is necessary so that copyright protection does not 'extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work,' as proscribed by the Copyright Act." (quoting 17 U.S.C. § 102(b))).

270. *Id.* at 1336–37; *see also* Pamela Samuelson, *Copyright and Freedom of Expression in Historical Perspective*, 10 J. INTELL. PROP. L. 319, 340 (2003). *See generally* Mark A. Lemley, *The Benefit of the Bargain*, 2023 WIS. L. REV. 237 (discussing enforcement of non-negotiated contracts that override default rules).

271. Computer Fraud and Abuse Act of 1986, Pub. L. 99-474, 100 Stat. 1213 (codified at 18 U.S.C. § 1030). This analysis of the CFAA draws on my previous co-authored work. Emine Özge Yildirim, Molly Shaffer Van Houweling, Ana Lazarova & Brigitte Vézina, *Freedom to Share: How Government's Data Sharing Policies Concerning Publicly Available Data Impact Academic Research and Journalism in the Public Interest* (Jan. 11, 2023) (manuscript at 9–10), https://ssrn.com/abstract=4484001 [https://perma.cc/A2HC-PEX3 (staff-uploaded archive)].

272. 18 U.S.C. § 1030(a)(2)(C).

273. *Id.* § 1030(e)(2)(B).

notwithstanding the fact that the data would be fully extractable as a matter of copyright law.[274]

The trend in recent cases has been toward a more limited reading of the CFAA, to prohibit access only to information that is behind a technological gate (e.g., password protection), not to publicly available information that is being used in a way the platform owner does not like. In *Van Buren v. United States*,[275] for example, the Supreme Court endorsed a "gates-up-or-down inquiry" under which authorization turns on whether someone "can or cannot access a computer system."[276] As Orin Kerr has pointed out, the Supreme Court did not make it entirely clear what counts as a "gate."[277] But mere admonitions posted on publicly available websites are probably not enough.[278]

In *hiQ Labs, Inc. v. LinkedIn Corp.*,[279] the Ninth Circuit observed that *Van Buren* "reinforces our conclusion that the concept of 'without authorization' does not apply to public websites."[280] The court went on to articulate the importance of this limitation in a way that evokes the freedom to extract: "giving companies like LinkedIn free rein to decide, on any basis, who can collect and use data . . . risks the possible creation of information monopolies that would disserve the public interest."[281]

In addition to these copyright-adjacent fields of law that impact the freedom to extract noncopyrightable material, it is instructive to explore how other fields of intellectual property address analogous situations involving potential obstacles to the extraction of unprotectable material. In IP contexts

274. *See, e.g.*, Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067–69 (9th Cir. 2016); Craigslist Inc. v. 3Taps Inc., 942 F. Supp. 2d 962, 970 n.8 (N.D. Cal. 2013); Sw. Airlines Co. v. Farechase, Inc., 318 F. Supp. 2d 435, 438–39 (N.D. Tex. 2004); Register.com, Inc. v. Verio, Inc., 126 F. Supp. 2d 238, 251 (S.D.N.Y. 2000), *aff'd as modified*, 356 F.3d 393 (2d Cir. 2004); *see also* Charles Duan, *Hacking Antitrust: Competition Policy and the Computer Fraud and Abuse Act*, 19 COLO. TECH. L.J. 313, 335–36 (2021); Christine D. Galbraith, *Access Denied: Improper Use of the Computer Fraud and Abuse Act to Control Information on Publicly Accessible Internet Websites*, 63 MD. L. REV. 320, 323 (2004); Mark A. Lemley, *Place and Cyberspace*, 91 CALIF. L. REV. 521, 528 (2003). *See generally* Jonathan Mayer, *Cybercrime Litigation*, 164 U. PA. L. REV. 1453 (2016) (discussing overbreadth of cybercrime liability); Orin S. Kerr, *Cybercrime's Scope: Interpreting "Access" and "Authorization" in Computer Misuse Statutes*, 78 N.Y.U. L. REV. 1596 (2003) (same).

275. 593 U.S. 374 (2021).

276. *Id.* at 376.

277. Orin S. Kerr, *Focusing the CFAA in* Van Buren, 2021 SUP. CT. REV. 155, 156 (2021).

278. As the Court explained, "[i]f the 'exceeds authorized access' clause criminalizes every violation of a computer-use policy, then millions of otherwise law-abiding citizens are criminals." *Van Buren*, 593 U.S. at 394. For example, "[m]any websites, services, and databases—which provide 'information' from 'protected computer[s],' § 1030(a)(2)(C)—authorize a user's access only upon his agreement to follow specified terms of service." *Id.* And so, "[i]f the 'exceeds authorized access' clause encompasses violations of circumstance-based access restrictions on employers' computers, it is difficult to see why it would not also encompass violations of such restrictions on website providers' computers." *Id.*

279. 31 F.4th 1180 (9th Cir. 2022).

280. *Id.* at 1199.

281. *Id.* at 1202.

outside of copyright, courts sometimes recognize that rights to seemingly protected material must yield to ensure that defendants have access to unprotected material. Two particularly interesting examples come from trademark law.

Trademark law generally recognizes that competitors should be free to use words and symbols in connection with their products and services where those words and symbols help to communicate the identity or characteristics of the products or services. At the same time, the law recognizes that words and symbols that describe a product can come, over time, to be associated with a particular source. This phenomenon of "acquired distinctiveness through secondary meaning" can produce "descriptive" trademarks—that is, words or symbols that both describe a product or service *and* serve as a source identifier for the mark owner.[282] When this happens, nonowners are prohibited from using the mark as a source identifier in a way likely to cause consumer confusion.[283] However, ownership rights for a descriptive trademark do not prohibit nonowners from using the word or symbol in its original, descriptive way.[284] They can still use it to accurately describe the characteristics of their own products or services.[285] In other words, they can *extract* the descriptive value of the word or symbol even though using it as a confusing source identifier is forbidden. This is the case even if some consumers might be confused by the use of a protected mark in this way.[286] These dual-purpose descriptive marks—protectable as source identifiers but subject to unauthorized use for descriptive purposes—are akin to copyrightable expression that may be freely used for functional, evidentiary, or other purposes designed to leverage the unprotectable aspects of the work.

Another analogous example from trademark law is the phenomenon of genericide. In general, terms or symbols that identify an entire category of product or service are deemed "generic" and cannot be protected as trademarks at all.[287] "Genericide" is the process by which a mark becomes generic, and thus unprotectable, over time.[288] This happens when a brand name that serves as a source identifier comes to signify the entire category of products in the minds of consumers.[289] This was the fate of marks like "cellophane," "escalator," and

282. Two Pesos, Inc. v. Taco Cabana, Inc., 505 U.S. 763, 769 (1992).

283. *See* 15 U.S.C. §§ 1114(1), 1125(a)(1)(A).

284. *See id.* § 1115(b)(4).

285. *See id.*

286. *See* KP Permanent Make-Up, Inc. v. Lasting Impression I, Inc., 543 U.S. 111, 121 (2004).

287. U.S. Pat. & Trademark Off. v. Booking.com B.V., 591 U.S. 549, 554 (2020); RESTATEMENT (THIRD) OF UNFAIR COMPETITION § 15 (AM. LAW INST. 1995).

288. *See, e.g.*, Elliott v. Google, Inc., 860 F.3d 1151, 1155–56 (9th Cir. 2017).

289. *See, e.g.*, Freecycle Network, Inc. v. Oey, 505 F.3d 898, 905 (9th Cir. 2007).

"murphy bed."[290] The phenomenon of genericide demonstrates how the needs of competitors—and the public—to use material subject to intellectual property protection can change over time in a way that alters the protectability of the material. I will return to this concept below, where I argue that copyright courts sometimes fail to adequately recognize how constraints on the freedom to extract can similarly change over time in a way that should alter the copyright analysis.

In this Part, I have explored how courts have applied copyright and related areas of law to vindicate the freedom to extract, even as technological developments have changed the environment in which that freedom is exercised. I have also noted a few outlying cases in which courts have taken positions that endanger that freedom. This exploration has been wide-ranging but not exhaustive. There are additional doctrines that protect the freedom to extract that I leave to be explored in future work. These include, for example, the first sale doctrine—which facilitates the circulation of lawful copies of works to new readers who might extract their insights. In the next Part, I explore threats to the freedom to extract and how copyright doctrine might be improved to withstand them.

II. IMPROVING THE LAW TO SAFEGUARD THE FREEDOM TO EXTRACT

The evolution of the freedom to extract explored in Part I demonstrates that technological developments can hinder or help extraction. The dissemination of software in object code, the embedding of facts in computerized databases, and the locking of unprotected elements behind technological protection measures all make extraction more difficult. Advances in computational power and techniques make extraction more powerful—and potentially more beneficial to society.[291] Courts generally have recognized both of these possibilities in a way that safeguards extraction from technological obfuscation and allows for the harnessing of technology that facilitates extraction. Contemporary developments—most notably in artificial intelligence—will test courts' ability to assess and address technology's impacts on extraction.[292] I will return to the issue of AI in Part III after first exploring

290. Elliott, 860 F.3d at 1156; Murphy Door Bed Co. v. Interior Sleep Sys., Inc., 874 F.2d 95, 101 (2d Cir. 1989).

291. *See* Sag, *Copyright and Copy-Reliant Technology*, *supra* note 47, at 1625 ("The legal status of actual copying for nonexpressive uses was not a burning issue before digital technology: there simply was no commercially relevant total literal copying directed towards a nonexpressive end . . . . However, digital technology and the increasing value of metadata have combined to make the legality of nonexpressive copying arguably the most significant issue in copyright law today.").

292. *See* Lemley & Casey, *supra* note 3, at 775–76 (describing machine learning systems as "robotic learners" that must necessarily copy protectable aspects of creative works in order to learn about the ideas embedded in those works).

more generally how current doctrine could be improved to better safeguard and harness extraction.

The improvements I suggest fall into three categories. First, courts could more consistently recognize how the necessity for, and difficulty of, extracting can change over time. This can mean that subject matter that was initially protectable can become unprotectable over time (and possibly vice versa). Second, where extractive use is at issue, that should influence every factor in the fair use analysis. Finally, regardless of the specific doctrinal question at issue, courts should consider the extent to which a given extractive use furthers the purpose of copyright to promote progress in human knowledge. This final proposed improvement is the most novel. It would require courts to consider not only whether a defendant's actions were designed to extract, but also whether they facilitate downstream extraction by others or, to the contrary, obfuscate unprotected elements of the defendant's own work. In other words, the law should consider whether a defendant's activity is *extractable* in addition to whether it is extractive.

A. *Recognizing How the Need to Extract Can Change over Time*

As the many examples described above make clear, the technological context in which copyrighted works are situated can have an impact on the ease or difficulty of exercising the freedom to extract. The doctrines that vindicate that right should be sensitive to that context and their applicability should therefore be flexible over time. But courts sometimes see elements of key extraction doctrines as one-shot, on/off determinations that lack this context-specificity and flexibility.

The question of whether the need to copy expression in order to exercise the freedom to extract can change over time has been a contentious one in the context of the merger doctrine. That doctrine could be improved by more consistently recognizing that merger can happen over time.

As discussed above, vindication of the freedom to extract is the essence of the merger doctrine. Courts consistently explain that merger applies to exclude expression from protection where the alternative would, as a practical matter, extend protection to ideas, methods, or facts. This is important because, without such a doctrine, copyright could operate to drastically limit the ability of subsequent authors to engage with the substance of preexisting works, thereby foreclosing both competition and communication. The focus of the analysis should therefore be on the constraints that a would-be user faces at the time of that engagement. The question "is whether other options practically exist under the circumstances."[293] For the merger doctrine to safeguard the freedom to

293. Lexmark Int'l, Inc. v. Static Control Components, Inc., 387 F.3d 522, 536 (6th Cir. 2004), *abrogated on other grounds by* eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388 (2006).

extract, those "circumstances" should include the circumstances facing the would-be extractor.

But even courts that express a fundamental understanding of the rationale for the merger doctrine sometimes apply the doctrine in ways that do not account for all relevant practical constraints that limit the alternatives for expressing ideas, methods, or facts. In particular, some courts refuse to apply merger to account for constraints that are faced by a defendant if those same constraints were not faced by the creator of the copyrighted work. This can happen when new constraints arise over time (e.g., because of changing technology standards or market conditions), or when the very act of creating the work in the first place imposes constraints on subsequent users (e.g., when the expression at issue is a lockout code that must be used in order to access unprotected aspects of the work).[294]

Consider, for example, the Federal Circuit's merger analysis in *Oracle America, Inc. v. Google Inc.*[295] The court quoted, with apparent approval, this language about the rationale for the merger doctrine: "Under the merger doctrine, courts will not protect a copyrighted work from infringement if the idea underlying the copyrighted work can be expressed in only one way, lest there be a monopoly on the underlying idea."[296] The focus on monopoly suggests a concern with competition, and thus with the constraints faced by the copyright owners' competitors—including the defendant who is accused of copying copyrighted expression. And yet the Federal Circuit refused to consider merger from the defendant's perspective, faulting the district court for "focusing its merger analysis on the options available to [defendant] Google at the time of copying."[297] The court insisted that "copyrightability and the scope of protectable activity are to be evaluated at the time of creation, not at the time of infringement" and that "[t]he focus is, therefore, on the options that were available to [plaintiff] Sun/Oracle at the time it created" the software interfaces

294. On analyzing merger from the defendant's perspective, see, for example, Menell, *Economic Analysis of Network Effects*, *supra* note 75, at 178; Samuelson, *Reconceptualizing Copyright's Merger Doctrine*, *supra* note 67, at 443–44; Joseph Gratz, *Merger as a Matter of Extrinsic Constraints*, 43 COLUM. J.L. & ARTS 403, 404–05 (2020); Peter S. Menell, *An Analysis of the Scope of Copyright Protection for Application Programs*, 41 STAN. L. REV. 1045, 1100–01 (1989); Stephen Preonas, Comment, *Mergercide, When Good Copyrights Go Bad: A Recommendation for a Market-Based, Defendant-Centric Approach to the Merger Doctrine in the Context of Compilations*, 11 INTELL. PROP. L. BULL. 89, 97–99 (2006); Timothy S. Teter, Note, *Merger and the Machines: An Analysis of the Pro-Compatibility Trend in Computer Software Copyright Cases*, 45 STAN. L. REV. 1061, 1075–87 (1993); *see also* Peter Lee, *The Evolution of Intellectual Infrastructure*, 83 WASH. L. REV. 39, 45 (2008) (explaining that "trademark and copyright law use social-feedback mechanisms to relax exclusive rights on applications that have become infrastructure").

295. 750 F.3d 1339, 1359–62 (Fed. Cir. 2014).

296. *Id.* at 1361 (quoting Satava v. Lowry, 323 F.3d 805, 812 n.5 (9th Cir. 2003)).

297. *Id.*

at issue in the case.[298] This way of applying the merger doctrine fails to safeguard the freedom to extract because it ignores the practical realities faced by the potential extractor and thus makes it possible for copyright owners to establish a "monopoly on the underlying idea" (or method, or fact). The Federal Circuit compounded the error by taking the same time-bound, one-shot approach to scènes à faire.[299]

When the case eventually reached the Supreme Court, this issue was raised in the briefs but not resolved.[300] The Court held Google's copying to be fair use without deciding the question of whether that copying in fact amounted to copyright infringement (or, instead, merely copying of merged and therefore unprotected elements).[301] But the Court's observations about the nature of the software interfaces at issue in the case seem to acknowledge the importance of the freedom to extract. In concluding that the second fair use factor (the "nature of the copyrighted work") favored Google, the Court observed the ways in which the code at issue (which the Federal Circuit said could have initially been written in many different ways) was "inextricably bound together with a general system, the division of computing tasks, that no one claims is a proper subject of copyright" and "inextricably bound up with the idea of organizing tasks . . . that is also not copyrightable" and "inextricably bound up with the use of specific commands known to programmers."[302] This entanglement of expression with an unprotectable system and unprotectable idea—and with the marketplace reality that programmers had learned *over time* to use specific commands—weighed in favor of fair use, which the Court explained "can play an important role in determining the lawful scope of a computer program copyright" and can "provid[e] a context-based check that can help to keep a copyright monopoly within its lawful bounds."[303] Thus the Court's opinion

298. *Id.*; *see also* U.S. COPYRIGHT OFF., SOFTWARE-ENABLED CONSUMER PRODUCTS 16 (2016) [hereinafter U.S. COPYRIGHT OFF., SOFTWARE-ENABLED] ("For both merger and *scènes à faire*, courts must focus on the options available to the author at the time a work is initially created, rather than the choices available to users after the fact.").

299. *Oracle*, 750 F.3d at 1363–64; *see also* Dun & Bradstreet Software Servs., Inc. v. Grace Consulting, Inc., 307 F.3d 197, 215 (3d Cir. 2002) (scènes à faire); Mitel, Inc. v. Iqtel, Inc., 124 F.3d 1366, 1375 (10th Cir. 1997) (same); U.S. COPYRIGHT OFF., SOFTWARE-ENABLED, *supra* note 298, at 16 (merger and scènes à faire).

300. Brief for the Petitioner at i, Google LLC v. Oracle Am., Inc., 593 U.S. 1 (2021) (No. 18-956); Brief for Respondent at i, Google LLC v. Oracle Am., Inc., 593 U.S. 1 (2021) (No. 18-956); *see also* Google LLC v. Oracle Am., Inc., 593 U.S. 1, 16 (2021) (noting that "[Google] asked [the Court] to review the Federal Circuit's determinations as to both copyrightability and fair use"); *id.* at 48 (Thomas, J., dissenting) (addressing and rejecting Google's merger argument).

301. *Google*, 593 U.S. at 20. Regarding the overlap between the merger and fair use doctrines, see Bracha & Golden, *supra* note 200, at 256–73 (discussing the "relationship between the rule that denies copyright protection to functional subject matter and the fair use doctrine" and concluding that "partial redundancy between § 102(b) limitations [including merger] and the fair use doctrine is beneficial").

302. *Google*, 593 U.S. at 27.

303. *Id.* at 22.

acknowledges the possibility that copying protected expression may sometimes be necessary to extract a work's unprotected elements, and that this necessity may arise because of market constraints that arise over time. It thus reiterates the importance of the freedom to extract. Unfortunately, because it did not decide the merger issue, the Court missed the opportunity to fully correct the Federal Circuit's endangerment of that right.[304]

Other courts have recognized the possibility of merger where constraints either result from the initial author's expressive choices or emerge over time. In *Veeck v. Southern Building Code Congress International, Inc.*,[305] the work at issue was a privately-authored building code that was later adopted as the official building code of two Texas towns.[306] The Fifth Circuit clearly held that these building codes were initially copyrightable.[307] But when the codes were later adopted by the towns, their expression became the only way to express the "fact" or "idea" of what the law was, and therefore the merger doctrine applied.[308] More recently, the district court in *International Code Council, Inc. v. UpCodes, Inc.*[309] faced similar facts and directly addressed the question of merger over time.[310] The court rejected the Federal Circuit's approach to the question in *Oracle*, explaining that "when assessing Defendants' alleged infringement, [the court] should consider whether previously copyrighted language *has become* essential to the expression of, or integrated with, a legal conception."[311] Looking to the governing law of the Second Circuit (and following the persuasive precedent of *Veeck*), the court held that it "need not ignore changes that transpire between a work's initial creation and its alleged infringement, particularly because the manner in which people use or rely on that work may fundamentally change in the interval."[312]

In *Lexmark*, the court applied a similar dynamic understanding to both merger and scènes à faire. In that case, the initial author's use of a lock-out code

304. *See* Mark A. Lemley & Pamela Samuelson, *Interfaces and Interoperability After* Google v. Oracle, 100 TEX. L. REV. 1, 1 (2021).

305. 293 F.3d 791 (5th Cir. 2002).

306. *Id.* at 791.

307. *Id.* at 802.

308. *Id.* The court concluded its merger analysis this way:

> We emphasize that in continuing to write and publish model building codes, SBCCI is creating copyrightable works of authorship. *When* those codes are enacted into law, however, *they become* to that extent "the law" of the governmental entities and may be reproduced or distributed as "the law" of those jurisdictions.

Id. (emphasis added). Regarding the importance of freedom to extract the law (and to reuse copyrighted works from which extraction of the law is difficult), see Grimmelmann, *Opinionated Primer*, *supra* note 3, at 27–28.

309. No. 17 CIV. 6261, 2020 WL 2750636 (S.D.N.Y. May 27, 2020).

310. *Id.* at *19.

311. *Id.* at *21 (emphasis added).

312. *Id.*

narrowed the range of alternatives available to *subsequent* authors seeking to create programs that were compatible with Lexmark's printers.[313] For that reason, the Sixth Circuit found that the lock-out code was unprotectable by copyright: "To the extent compatibility requires that a particular code sequence be included in the component device to permit its use, the merger and scènes à faire doctrines generally preclude the code sequence from obtaining copyright protection."[314] This approach—considering compatibility requirements from the defendant's perspective and thus acknowledging that they can change over time—reflects the rationale underlying the merger and scènes à faire doctrines and helps to safeguard the freedom to extract.

This dynamic and context-sensitive application of the merger and scènes à faire doctrines makes sense in light of the rationales animating the doctrines.[315] These doctrines are bulwarks of the idea/expression distinction, which preserves breathing space for competition and communication. The doctrines should therefore be sensitive to the cultural, economic, and even legal importance of the material at issue,[316] all of which can change over time.

Deploying merger or scènes à faire to reinforce the idea/expression distinction when the constraints on a defendant have changed over time is consistent with the First Amendment, as well as copyright policy. The Supreme Court has repeatedly explained that copyright is consistent with the First Amendment, despite the constraints it imposes on speech, because copyright's "built-in First Amendment accommodations" (specifically, fair use and the idea/expression distinction) ensure that defendants can express themselves using alternatives to (or fair uses of) copyright owners' protected expression.[317] For First Amendment purposes, the adequacy of a copyright defendant's

313. Lexmark Int'l, Inc. v. Static Control Components, Inc., 387 F.3d 522, 531 (6th Cir. 2004), *abrogated on other grounds by* eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388 (2006).

314. *Id.* at 536.

315. Regarding merger, see *supra* notes 294–95 and accompanying text. Regarding scènes à faire, see Bracha & Golden, *supra* note 200, at 267; Peter Lee & Madhavi Sunder, *The Law of Look and Feel*, 90 S. CAL. L. REV. 529, 581 (2017); Mala Chatterjee, *Intellectual Property, Independent Creation, and the Lockean Commons*, 12 U.C. IRVINE L. REV. 747, 782–83 (2022); Kim, *supra* note 73, at 147–48.

316. *See* Herbert Rosenthal Jewelry Corp. v. Kalpakian, 446 F.2d 738, 742 (9th Cir. 1971) (explaining in a case decided on merger grounds that "[w]hat is basically at stake is the extent of the copyright owner's monopoly—from how large an area of activity did Congress intend to allow the copyright owner to exclude others?").

317. Eldred v. Ashcroft, 537 U.S. 186, 219–20 (2003); *see also* Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 560 (1985) (explaining that First Amendment protections are "embodied in the Copyright Act's distinction between copyrightable expression and uncopyrightable facts and ideas," and in the "latitude for scholarship and comment" safeguarded by the fair use defense); *id.* at 556 ("The Second Circuit noted, correctly, that copyright's 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, Publishers, Inc. v. Nation Enters., 723 F.2d 195, 203 (1983))); United States v. Bodin, 375 F. Supp. 1265, 1267 (W.D. Okla. 1974); Sid & Marty Krofft Television Prods., Inc. v. McDonald's Corp., 562 F.2d 1157, 1170–71 (9th Cir. 1977).

expressive alternatives should be assessed from the perspective of defendants, at the time when they want to express themselves.[318] And so, to operate effectively as First Amendment safeguards, fair use and the idea/expression dichotomy need to operate from this perspective as well. And merger and scènes à faire need to do so too, insofar as they are essential to ensuring that copyright protection does not, as a practical matter, extend to ideas and other unprotected elements.[319]

So far in this section, I have addressed ways in which the nature of an element within a work might change over time, thus changing the necessity of extracting it. Original expression might become the law over time, for example, or it may become an interoperability key, or it may become stock within a genre. All of those developments can trigger the need to vindicate the freedom to extract. Merger and scènes à faire should adjust over time to provide that vindication. Some courts recognize this, but others do not, thus endangering the freedom to extract.

There is another way in which the context for the freedom to extract can change over time: the *difficulty* of extraction can change. We saw this in a general way in the discussion of the emergence of computer software as a type of protectable work. With the rise of works that were distributed only in non-human-readable form, extracting unprotected elements became more difficult. In the face of this technological development, some courts started to recognize the permissibility of copying works in their entireties where necessary to vindicate the freedom to extract.

This phenomenon can also happen at the level of an individual work. Imagine, for example, a work that is available only in a single copy in an

318. Regarding the relevance of adequate alternatives in the context of First Amendment conflicts with tangible and intellectual property rights apart from copyright, see, for example, Lloyd Corp., Ltd. v. Tanner, 407 U.S. 551, 567 (1972); Dall. Cowboys Cheerleaders, Inc. v. Pussycat Cinema, Ltd., 604 F.2d 200, 206 (2d Cir. 1979); Mut. of Omaha Ins. Co. v. Novak, 836 F.2d 397, 402 (8th Cir. 1987). For discussion of how courts analyze First Amendment defenses in copyright cases in light of the availability of other means to express ideas, see, for example, Pamela Samuelson, *Reviving* Zacchini*: Analyzing First Amendment Defenses in Right of Publicity and Copyright Cases*, 57 TUL. L. REV. 836, 878–83 (1983). The existence of alternative avenues of communication is relevant more generally when First Amendment challenges are brought against content-neutral speech restrictions. *See* Clark v. Cmty. for Creative Non-Violence, 468 U.S. 288, 293 (1984) ("We have often noted that [reasonable time, place, or manner restrictions] are valid provided that they are justified without reference to the content of the regulated speech, that they are narrowly tailored to serve a significant governmental interest, and that they leave open ample alternative channels for communication of the information."). For a discussion of whether copyright law should be analyzed as a content-neutral or content-based speech restriction, see Neil Weinstock Netanel, *Locating Copyright Within the First Amendment Skein*, 54 STAN. L. REV. 1, 48–59 (2001).

319. Regarding copyright's First Amendment safeguards, and how free speech is endangered when they are narrowed, see generally Mark A. Lemley & Rebecca Tushnet, *First Amendment Neglect in Supreme Court Intellectual Property Cases*, 2023 SUP. CT. REV. 85 (2024).

ephemeral medium.[320] The ability to extract unprotected elements over time may change as that copy degrades. In such a case, it may become necessary to copy the entire work in order to preserve the opportunity to extract unprotected elements from it before it disappears. Although I know of no court to have done so, one can imagine merger being deployed in such a circumstance. It would be literally true in such a case that "protection of the expression would effectively accord protection to the idea itself,"[321] which is the essence of merger. It seems more likely, however, that such a scenario would be addressed under fair use, to which I now turn.[322]

B. *Extraction and the Fair Use Factors*

The fair use doctrine could better safeguard the freedom to extract by adjusting analysis of each of the four of the fair use factors.

1. Factor One

With regard to the first factor, "the purpose and character of the use,"[323] courts should more clearly and consistently recognize extractive use as a valid purpose.[324]

Although extractive use overlaps with some purposes courts already consistently favor in their fair use analysis—most notably "transformative" use—this is not always the case.[325] In its most recent fair use decision, *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, the Supreme Court described transformativeness in terms of the defendant's purpose being different from the copyright owner's purpose.[326] Extraction is a purpose that is typically different from the copyright owner's purpose; at the same time, extraction can be conducted in the context of an activity that is generally for the *same* purpose. In *Sega v. Accolade*, for example, the defendant's overarching

320. For scholarship exploring the phenomenon of works becoming entirely unavailable, see, for example, Mark A. Lemley, *Disappearing Content*, 101 B.U. L. REV. 1255, 1257 (2021); R. Anthony Reese, *Super Bowl I, Jazz Radio, and* The Glass Menagerie*: Copyright, Preservation, and Private Copies*, 51 AKRON L. REV. 1025, 1027–29 (2017); Anthony Reese, *What Copyright Owes the Future*, 50 HOUS. L. REV. 287, 287–89 (2012); Eva E. Subotnik, *Copyright and the Living Dead?: Succession Law and the Postmortem Term*, 29 HARV. J.L. & TECH. 77, 81–82 (2015).

321. Kregos v. Associated Press, 937 F.2d 700, 705 (2d Cir. 1991).

322. In the context of library preservation, this scenario could also be addressed by 17 U.S.C. § 108(b), (c), (h) (providing for preservation of works and replacement of damaged copies under specified circumstances).

323. 17 U.S.C. § 107(1).

324. *See* Lemley & Casey, *supra* note 3, at 750, 776 (arguing that use of copyrighted works for artificial intelligence training should be presumptively fair under the first factor if the purpose "is not to obtain or incorporate the copyrightable elements of a work but to access, learn, and use the unprotectable parts of the work").

325. *Cf.* Sag, *Orphan Works*, *supra* note 47, at 1535–37 (discussing the incomplete overlap between non-expressive use and transformative use).

326. *See* 598 U.S. 508, 529 (2023).

purpose was to make successful video games, just like the copyright owner.[327] A cookbook author who extracts unprotectable recipes from existing cookbooks has the same overarching purpose as the creators of the originals did.[328]

As examples like *Sega* demonstrate, courts sometimes squeeze extractive use into the "transformative use" box, thus favoring fair use. But the awkward fit between extraction and transformation makes this an unreliable technique for vindicating the freedom to extract. By clearly articulating extractive use as a valid purpose, courts might more consistently recognize that these seemingly non-transformative activities nonetheless vindicate an important pro-copyright goal of keeping fundamental elements in the public domain. Furthermore, recognizing a distinct category of extractive use will set the stage for recognizing the relevance of extraction to all of the other fair use factors—just as transformativeness tends to influence the entire fair use analysis today.[329]

There are several examples of courts focusing on transformation without also recognizing the value of extraction. These include cases in which photographs were arguably used in an effort to extract the factual information they conveyed.

In *Sedlik v. Drachenberg*,[330] photographer Jeffrey Sedlik alleged infringement based on the use of his photographic portrait as a reference for a tattoo depicting Miles Davis.[331] In considering tattoo artist Kat Von D's fair use argument, the court gave little attention to the possibility that she used the portrait for the valid purpose of extracting the unprotectable facts of Davis's appearance.[332] The court concluded that the tattoo did not "require borrowing from the original" based on Kat Von D's testimony that she would have "just

327. Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1522 (9th Cir. 1992); *see also* Ard, *supra* note 51 (manuscript at 41) (making a similar point about the overlap between the plaintiff's and defendant's purposes in *Sega*).

328. Although this example would not seem to justify a fair use analysis if in fact only unprotectable facts were copied, thus failing to establish a prima facie case of infringement, fair use could be determinative if some incidental expression was copied. Furthermore, courts often assume infringement and proceed directly to a fair use analysis, as the Supreme Court did in *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1 (2021).

329. Regarding the role of transformativeness in fair use analysis, see, for example, Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions Updated, 1978-2019*, 10 NYU J. INTELL. PROP. & ENT. L. 1, 25–28 (2020); Clark D. Asay, Arielle Sloan & Dean Sobczak, *Is Transformative Use Eating the World?*, 61 B.C. L. REV. 905, 922 (2020); Jiarui Liu, *An Empirical Study of Transformative Use in Copyright Law*, 22 STAN. TECH. L. REV. 163, 167 (2019); Matthew Sag, *Predicting Fair Use*, 73 OHIO ST. L.J. 47, 55–61 (2012) [hereinafter Sag, *Predicting Fair Use*]; Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978-2005*, 156 U. PA. L. REV. 549, 603–09 (2008) [hereinafter Beebe, *An Empirical Study, 1978-2005*]; R. Anthony Reese, *Transformativeness and the Derivative Work Right*, 31 COLUM. J.L. & ARTS 467, 485 (2008).

330. No. CV 21-1102 DSF (MRWx), 2023 WL 6787447 (C.D. Cal. Oct 10, 2023).

331. *Id.* at *1.

332. *Id.* at *4 (denying summary judgment to both parties due to triable issues of fact with regard to the first and fourth fair use factors), *appeal docketed*, No. 24-3367 (9th Cir. May 29, 2024).

used another image" if the plaintiff's portrait had not existed.[333] But the fact that another image of Miles Davis would have been interchangeable with the plaintiff's suggests that Kat Von D was not using the portrait for its unique expressive contributions, but rather as one of many possible sources of the facts of Davis's appearance. That is, she was making extractive use that should have been recognized as such for purposes of the first fair use factor.

In *McGucken v. Pub Ocean Ltd.*,[334] the Ninth Circuit held that the first fair use factor weighed against the unauthorized publication of a photograph of an "ephemeral lake" that formed after a rare weather event in Death Valley.[335] The fact that the photograph conveyed factual information and was used by the defendant to depict a rare natural phenomenon did not count in favor of fair use because that purpose was the same as the copyright owner's and therefore not transformative.[336] By focusing only on transformativeness, the court failed to see the possibility that the use was validly *extractive*, insofar as it was necessary to fully convey the facts of the rare event.

McGucken relied heavily on *Monge v. Maya Magazines, Inc.*, mentioned above, in which the defendant published previously unpublished celebrity wedding photos.[337] The court held that the defendant's use was not transformative because it did not alter the aesthetics of the original.[338] The fact that the defendant's purpose "was to expose the couple's secret wedding"—that is, to expose a previously unknown historical fact—did not amount to transformation and did not, in the majority's view, weigh in favor of fair use.[339] Again, the court missed the opportunity to recognize that a use that is not transformative may nonetheless be validly *extractive* if its purpose is to communicate the facts embedded in a copyrighted work.

The Supreme Court similarly gave short shrift to extraction in *Warhol v. Goldsmith*, where the defendant's purpose, "to depict Prince in magazine stories about Prince," was both identical to the plaintiff's purpose *and* extractive of the

333. *Id.* at *4.

334. 42 F.4th 1149 (9th Cir. 2022).

335. *Id.* at 1161.

336. *Id.* at 1158 ("Under our case law, a work that conveys factual information does not transform a copyrighted work by using it as a 'clear, visual recording' of the infringing work's subject." (quoting Monge v. Maya Mags., 688 F.3d 1164, 1174 (9th Cir. 2012))); *id.* at 1153 ("Pub Ocean's use was in no way transformative—the article used McGucken's photos to depict the ephemeral lake, which was exactly the purpose for which they were taken and exactly the function for which the photos had been licensed to other websites.").

337. *Monge*, 688 F.3d.at 1169–70.

338. *Id.* at 1176.

339. *Id.*; *see also* L.A. News Serv. v. Reuters Television Int'l, Ltd., 149 F.3d 987, 993 (9th Cir. 1998), *as amended on denial of reh'g and reh'g en banc* (Aug. 25, 1998). *But see Monge*, 688 F.3d at 1185–91 (Smith, J., dissenting) (arguing that the defendant's "commentary, editing, and arrangement of the photos added to, and ultimately changed, the original character of the images by advancing them as the basis of an exposé").

fact of Prince's actual appearance.[340] In focusing on the question of transformativeness (and, specifically, on the question of whether the defendant's purpose was *distinct* from the plaintiff's), the Court failed to acknowledge the importance of extraction.[341] The *Warhol* Court did, however, recognize one circumstance in which the first factor may weigh in favor of a defendant whose purpose is the same as that of the original work—where the defendant's work provides a commentary on the plaintiff's work:

> Parody, for example, needs to "mimic an original to make its point." Similarly, other commentary or criticism that targets an original work may have compelling reason to "conjure up" the original by borrowing from it. An independent justification like this is particularly relevant to assessing fair use where an original work and copying use share the same or highly similar purposes . . . .[342]

Thus, a parodic work that generally aims to entertain its audience through pop music can be fair even if it incorporates a copyrighted work that has the same general purpose. These of course were the facts of *Campbell v. Acuff Rose Music, Inc.*,[343] the earlier Supreme Court case on which the *Warhol* Court based its discussion of parody.[344] The need to conjure up the preexisting work to comment effectively upon it is a valid justification for copying. Neither *Campbell* nor *Warhol* held that this is the *only* valid justification for copying, even where the defendant shares the same general purpose as the copyrighted work. But they did not explore what other justifications might be. And, unfortunately, the *Warhol* Court failed to grapple with the extraction of unprotected facts as a potential justification despite the fact that the allegedly infringing work at issue arguably extracted the facts of Prince's actual appearance from the underlying work.[345]

The *Warhol* Court's inattention to the possibility of extractive use as a valid justification for copying was replicated and exacerbated by a panel of the

340. *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 526 (2023).

341. *Id.* at 525–33.

342. *Id.* at 511 (quoting Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 588 (1993)).

343. 510 U.S. 569 (1993).

344. *Warhol*, 598 U.S. at 510–11.

345. A version of this argument was presented to the Court. *See* Brief of Amici Curiae Copyright Law Professors in Support of Petitioner at 21, Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023) (No. 21-869), 2022 WL 2288299, at *21 ("Goldsmith has no copyright interest in what Price looked like, how he kept his facial hair, or the angle of his chin relative to his neck."); *see also* Mark A. Lemley, *How Generative AI Turns Copyright Upside Down*, 25 COLUM. SCI. & TECH. L. REV. 190, 210 (2024) (observing that "[t]he original theory for copyrighting photographs was that the artist was entitled to very narrow protection limited to the creative choices early photographers made in lighting and setting up their shots," but "[t]oday courts have forgotten those constraints, and regularly declare that photographs are entitled to strong protection even when virtually nothing about the photograph can be traced to authorial choices, as in *Warhol*" (citations omitted)).

Tenth Circuit in *Whyte Monkee Productions, LLC v. Netflix, Inc.*[346] In an opinion that has since been vacated pending rehearing en banc, the court considered a copyright claim based on the unauthorized incorporation of short video clips into the Netflix series "Tiger King."[347] The district court had granted the defendants' motion for summary judgment on the basis of fair use, but the Tenth Circuit reversed and remanded.[348] The appellate opinion took issue, in particular, with the district court's conclusion that the first fair use factor favored the defendants.[349] The Tenth Circuit held that the "purpose and character of the use" factor "strongly weighs in Plaintiffs' favor in light of the Supreme Court's recent guidance in *Warhol*."[350] The court agreed with the plaintiffs' argument that Netflix's use of the footage "is as commercial as it gets and is not transformative because the use makes no commentary upon the work itself."[351] It made no difference to the court that Netflix claimed the footage was necessary in order to communicate facts about Joe Exotic, the subject of its series—just as it seemed to make no difference to the *Warhol* Court that Andy Warhol had the purpose of depicting the facts of Prince's appearance. The court acknowledged that "Defendants used [the video] to illustrate Mr. Exotic's purported megalomania . . . . By doing so, Defendants were providing a historical reference point in Mr. Exotic's life and commenting on Mr. Exotic's showmanship."[352] But the court cited *Warhol* in holding that this extraction of a "historical reference point" was insufficient to favor fair use where the use did not comment on the copyrighted work.[353] Where the *Warhol* Court failed to consider the possibility that extraction of facts might serve as a justification for copying, the Tenth Circuit here considered the possibility and rejected it.

Oddly, the Tenth Circuit did seem to acknowledge extraction as a valid purpose in its consideration of another fair use factor. In assessing "the amount and substantiality of the portion used in relation to the copyrighted work as a whole,"[354] the court held that "Defendants used no more of [the video] than necessary, and what they did use was reasonable in light of providing historical reference points of Mr. Exotic's life, commenting on Mr. Exotic's showmanship, and creating a captivating viewing experience that would bring his story to life."[355]

346. 97 F.4th 699 (10th Cir.), *reh'g granted and opinion vacated*, 101 F.4th 787 (10th Cir. 2024).

347. *Id.* at 705.

348. *Id.* at 705–06.

349. *Id.* at 706.

350. *Id.* at 713.

351. *Id.*

352. *Id.* at 714.

353. *Id.* at 714–15 (observing that "in *Warhol*, Andy Warhol himself targeted a character—the artist, Prince—but the Court determined that his work was not sufficiently transformative in part because Mr. Warhol did not target the original work—*viz.*, Lynn Goldsmith's photograph of Prince").

354. 17 U.S.C. § 107(3).

355. *Whyte Monkee*, 97 F.4th at 719–20.

Where use of a copyrighted work is necessary to "provid[e] historical reference points," it arguably should be excused on the basis of merger without even getting to the question of fair use. But if such a case is decided on the basis of fair use, courts should recognize that the extractive purpose favors fair use whether or not that use can fairly be characterized as "transformative." The Supreme Court's failure to recognize this in *Warhol* is reverberating in the lower courts to the detriment of the freedom to extract.

The fair use analysis was more subtle in *Elvis Presley Enterprises, Inc. v. Passport Video*, mentioned above. There, the Ninth Circuit observed that the defendant's use of many television clips was "transformative because they are cited as historical reference points in the life of a remarkable entertainer."[356] The holding that the use was transformative is fairly debatable. The substance of the clips themselves was not significantly altered, and they were used to entertain their audience just as the originals were.[357] But the court could have reached a similar conclusion about the first factor by accurately characterizing the defendant's use as *extractive*, insofar as its purpose was to communicate the historical facts of the performances. Understanding the case in this light also shows that extractive use is not a get-out-of-jail-free card. Where the defendant's use went beyond that justified by extraction, the purpose was no longer valid: "[M]any of the film clips seem to be used in excess of this benign purpose, and instead are simply rebroadcast for entertainment purposes that Plaintiffs rightfully own."[358] Thus the district court's conclusion that the first factor weighed in favor of the plaintiffs was not clearly erroneous.

Not every extractive use will be fair, nor even result in the first factor weighing in the defendant's favor. But by focusing on the possibility of extractive use as a valid alternative (or complement) to transformative use, courts can avoid condemning behavior that is well-calibrated to copy elements that should rightly be in the public domain.

2. Factor Two

The second fair use factor, "the nature of the copyrighted work,"[359] should be shaped by recognition of the freedom to extract in two ways: by determining which aspect of a multifaceted work the defendant is using, and by considering how difficult it would be to cleanly extract unprotected elements from the copyrighted work.

356. Elvis Presley Enters., Inc. v. Passport Video, 349 F.3d 622, 629 (9th Cir. 2003), *overruled on other grounds by* eBay, Inc. v. MercExchange, L.L.C., 547 U.S. 388 (2006); *see also* Am. Inst. of Physics v. Winstead PC, No. 12-CV-1230-M, 2013 WL 6242843, at *5–7 (N.D. Tex. Dec. 3, 2013).

357. *Elvis*, 349 F.3d at 629.

358. *Id.*

359. 17 U.S.C. § 107(2).

First, courts should recognize that the nature of a copyrighted work can be multifaceted and they should focus on the aspect the defendant is exploiting. An elaborate password might be both a sonnet and a lock-out code, for example.[360] What should matter for purposes of fair use is what aspect the defendant's use is targeting. Where the defendant is extracting unprotected elements—for example, copying the sonnet in order to access software functionality—those elements should be the aspect of the "nature of the work" on which the analysis focuses.[361]

This type of subtle analysis of the second factor is rare. Although the Supreme Court breathed some new life into this neglected factor, starting off its fair use analysis with the second factor in *Google v. Oracle*,[362] the "nature of the copyrighted work" factor generally receives little attention and has little impact on the fair use analysis.[363] Courts typically give superficial attention to only two variables[364]—whether the copyrighted work is primarily factual versus creative[365] and whether it is published or unpublished.[366]

A rare example of the proposed approach in action is *Diamond v. Am-Law Publishing Corp.*,[367] in which the Second Circuit considered a copyright claim based on a magazine's publication of excerpts of a letter to the editor.[368] The letter addressed an earlier news story in the magazine related to a legal dispute involving the letter writer.[369] The letter writer had specified that the letter could

360. *See* Lexmark Int'l, Inc. v. Static Control Components, Inc., 387 F.3d 522, 544 (6th Cir. 2004), *abrogated on other grounds by eBay Inc.*, 547 U.S. 388.

361. The Supreme Court arguably took this approach in *Google v. Oracle*. *See* 4 MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 13F.06 (2023) ("[W]hat mattered for factor two [in *Google v. Oracle*] was not the nature and protectability of the work as a whole, but rather the nature and protectability of those portions copied.").

362. Google LLC v. Oracle Am., Inc., 593 U.S. 1, 26 (2021). For renewed attention to the second factor, see, for example, Trevor G. Reed, *Fair Use as Cultural Appropriation*, 109 CALIF. L. REV. 1373, 1421–42 (2021); Pamela Samuelson & Clark D. Asay, *Saving Software's Fair Use Future*, 31 HARV. J.L. & TECH. 535, 558 (2018); Clark D. Asay, *Software's Copyright Anticommons*, 66 EMORY L.J. 265, 317 (2017); Jennifer M. Urban, *How Fair Use Can Help Solve the Orphan Works Problem*, 27 BERKELEY TECH. L.J. 1379, 1392–402 (2012).

363. *See, e.g.*, PATRY, *supra* note 3, § 4.1; Robert Kasunic, *Is That All There Is? Reflections on the Nature of the Second Fair Use Factor*, 31 COLUM. J.L. & ARTS 529, 532, 544–47 (2008); Reed, *supra* note 362, at 1406; Leval, *supra* note 45, at 1116.

364. *See* Sag, *Predicting Fair Use*, *supra* note 329, at 61; Beebe, *An Empirical Study, 1978-2005*, *supra* note 329, at 610.

365. *See, e.g.*, Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 586 (1994); Stewart v. Abend, 495 U.S. 207, 237–38 (1990); Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 563 (1985).

366. *See, e.g.*, *Harper & Row*, 471 U.S. at 564. *But cf.* 17 U.S.C. § 107 ("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.").

367. 745 F.2d 142 (2d Cir. 1984), *abrogated on other grounds by* Fogerty v. Fantasy, Inc., 510 U.S. 517 (1994).

368. *Id.* at 144–46.

369. *Id.* at 144–45.

be published "only in its entirety."[370] And so the writer alleged that the publication of mere excerpts infringed his copyright.[371] The court rejected the copyright infringement claim on fair use grounds.[372] After explaining that the defendant's use of the letter amounted to news reporting regarding disputed questions of fact in connection with the original magazine story, the court continued to focus on the defendant's purpose in assessing the nature of the work, observing that the "use to which [the letter] was put was properly *informational*" and, therefore, "our discussion of factor (1) is thus dispositive as to (2)."[373] That is, because the defendant was using the copyrighted work in order to extract facts from it, the factor two analysis should also focus on the factual aspects of the work.

In other cases, courts of appeals have not used extraction to favor the defendant in the second factor analysis, but have at least used it to diffuse the factor in circumstances in which it might otherwise have favored the plaintiff. For example, in *Bill Graham Archives v. Dorling Kindersley Ltd.*,[374] the Second Circuit "agree[d] with the district court that the creative nature of artistic images typically weighs in favor of the copyright holder."[375] But where the defendant had used the creative Grateful Dead posters as historical artifacts, aiming to extract their documentary significance, the factor did not weigh significantly in the plaintiff's favor: "[E]ven though [the plaintiff]'s images are creative works, which are a core concern of copyright protection, the second factor has limited weight in our analysis because the purpose of [the defendant]'s use was to emphasize the images' historical rather than creative value."[376]

Similarly, in *Bouchat v. Baltimore Ravens Ltd. Partnership*,[377] the Fourth Circuit explained, in connection with use of a copyrighted logo in videos about football history, that "'if the disputed use of the copyrighted work is not related to its mode of expression but rather to its historical facts, then the creative nature of the work' matters much less than it otherwise would."[378]

Although they tend toward equivocation as opposed to full-throated endorsement of the freedom to extract, *Diamond*, *Bill Graham*, and *Bouchat* at least gesture toward an approach to the second factor that helpfully focuses on

370. *Id.* at 146.

371. *Id.* at 148.

372. *Id.*

373. *Id.* (emphasis added).

374. 448 F.3d 605 (2d Cir. 2006).

375. *Id.* at 612.

376. *Id.* at 612–13; *see also* Blanch v. Koons, 467 F.3d 244, 257 (2d Cir. 2006); A.V. *ex rel.* Vanderhye v. iParadigms, LLC, 562 F.3d 630, 640–42 (4th Cir. 2009); Bond v. Blum, 317 F.3d 385, 395–96 (4th Cir. 2003).

377. 737 F.3d 932 (4th Cir. 2013).

378. *Id.* at 943 (quoting *iParadigms*, 562 F.3d at 640).

those aspects of the copyrighted work the defendant was trying to exploit. More often, courts analyze the second factor by looking at the nature of the copyrighted work in isolation from the circumstances of the case.

For example, in *Wright v. Warner Books, Inc.*,[379] the Second Circuit insisted that "whether the infringer paraphrased or copied, whether he borrowed fact or expression, or whether his use implicates the author's privacy interests or not . . . have no bearing on factor two. Factor two focuses solely on the nature of the *copyrighted* work."[380] In *Hustler Magazine Inc. v. Moral Majority Inc.*,[381] the Ninth Circuit similarly rejected the idea that the nature of the work should be understood in light of the use to which the defendant put it: "The district court discounted the significance of the work's creative nature . . . because the defendants did not use the parody for its creative value. There is nothing in the statute, case law, or legislative history to support the district court's approach."[382] To the contrary, copyright law's long-standing recognition of the importance of the freedom to extract strongly supports focusing on the aspect of a work's nature that the defendant is exploiting under the particular circumstances of the case.

The second way in which factor two should be shaped by the freedom to extract is that courts should consider—as an aspect of a copyrighted work's "nature"—how difficult it would be to cleanly extract unprotected elements from the work. Where there is something about the work or the format in which it is disseminated that makes extraction difficult without copying protected expression, that aspect of the "nature of the copyrighted work" should count in favor of fair use.[383]

This was the Ninth Circuit's approach in both *Sega* and *Sony*, discussed above. In *Sega*, the court held that the second factor favored the defendant where "video game programs contain[ed] unprotected aspects that [could not] be examined without copying."[384] In *Sony*, the court held that the second factor "strongly favor[ed]" the defendant where copying of the work was necessary to access its unprotected elements.[385] The Federal Circuit also used this approach in *Atari Games Corp. v. Nintendo of America, Inc.*, explaining that "[w]hen the nature of a work requires intermediate copying to understand the ideas and processes in a copyrighted work, that nature supports a fair use for intermediate

379. 953 F.2d 731 (2d Cir. 1991).

380. *Id.* at 738.

381. 796 F.2d 1148 (2d Cir. 1986).

382. *Id.* at 1154.

383. *See* Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 563 (1985) ("The extent to which one must permit expressive language to be copied, in order to assure dissemination of the underlying facts, will . . . vary from case to case." (quoting Robert A. Gorman, *Fact or Fancy? The Implications for Copyright*, 29 J. COPYRIGHT SOC'Y U.S.A. 560, 563 (1982))).

384. Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1526 (9th Cir. 1992).

385. Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596, 603–605 (9th Cir. 2000).

copying."[386] Similarly, in *Consumers Union of United States, Inc. v. General Signal Corp.*,[387] the Second Circuit observed, in the context of factor two, that "[w]here an evaluation or description is being made, copying the exact words may be the only valid way precisely to report the evaluation."[388]

3. Factor Three

Recognition that a use is extractive should also impact the third fair use factor, "the amount and substantiality of the portion used in relation to the copyrighted work as a whole."[389] In extractive use cases, analysis of this factor should consider whether what was used was reasonable in light of the defendant's extractive purpose.

In theory, extractive use could involve use of only the unprotected "portion" of the original. But, as discussed repeatedly above, extraction of unprotected elements often involves incidental use of protected material. Courts considering extractive use cases should look to whether the defendant's use of elements beyond the unprotected elements it was aiming to extract was excessive in light of the difficulty of performing a perfect extraction.[390]

Especially in circumstances involving automated analysis of digitized works, it can be reasonably necessary to reproduce the entire copyrighted work in order to extract the unprotected facts, ideas, and methods within it.[391] The Second Circuit's analysis of the third factor in *Authors Guild v. Google* is consistent with this approach. The court explained why it was reasonable for Google to copy entire books in order to extract facts about those books: "If Google copied less than the totality of the originals, its search function could not advise searchers reliably whether their searched term appears in a book (or how many times)."[392]

Because courts have not fully recognized the concept of extractive use, they sometimes miss its relevance to the third factor. They fail to consider, for example, whether the amount and substantiality of the portion used was necessary in light of the difficulty of extracting unprotected elements. In *Warhol*

386. 975 F.2d 832, 843 (Fed. Cir. 1992).

387. 724 F.2d 1044 (2d Cir. 1983), *abrogated on other grounds* by Home Box Off., Inc. v. Showtime, Inc., 832 F.2d 1311 (2d Cir. 1987).

388. *Id.* at 1049–50 (citing Morrissey v. Procter & Gamble Co., 379 F.2d 675, 678–79 (1st Cir. 1967)).

389. 17 U.S.C. § 107(3).

390. *See* U.S. COPYRIGHT OFF., SECTION 1201 RULEMAKING, *supra* note 154, at 111 ("With respect to the third fair use factor, the amount and substantiality of the portion used in relation to the copyrighted work as a whole, copying the entire work [in connection with text and data mining] is likely reasonable in light of the purpose of the copying: obtaining data about the copyrighted works.").

391. *See* Sag, *The New Legal Landscape*, *supra* note 3, at 326 ("[M]aking complete digital copies of copyrighted works for text mining and similar non-expressive uses should be viewed as qualitatively insignificant under the third factor.").

392. Authors Guild v. Google, Inc., 804 F.3d 202, 221 (2d Cir. 2015).

v. Goldsmith, for example, the Second Circuit recognized that photographs capture unprotectable facts, including the appearance of the person who is the subject of a portrait.[393] But the court then went on to conclude that "while Goldsmith has no monopoly on Prince's face, the law grants her a broad monopoly on its image as it appears in her photographs of him."[394] And instead of recognizing that this "broad monopoly" should yield where necessary to permit extraction of the facts of Prince's appearance, the court concluded that copying a photograph is unreasonable under factor three even when the defendant has stripped away the identifiable creative elements:

> [W]here, as here, the secondary user has used the photograph itself, rather than, for example, a similar photograph, the photograph's specific depiction of its subject cannot be neatly reduced to discrete qualities such as contrast, shading, and depth of field that can be stripped away, taking the image's entitlement to copyright protection along with it.[395]

This analysis is backwards (but went uncorrected by the Supreme Court, which only reviewed the Second Circuit's analysis of the first factor).[396] In order to ensure that facts remain in the public domain, courts should recognize that in extractive use cases the reasonableness of the portion used should be calculated based in part on the difficulty of extracting unprotected elements. Copying even substantial portions of protected elements can be reasonable where necessary to extract unprotected elements. This of course is the essence of the merger doctrine, but it should also impact the fair use analysis in cases of extractive use.[397]

4. Factor Four

The final fair use factor is "the effect of the use upon the potential market for or value of the copyrighted work."[398] Here, courts can better vindicate the freedom to extract by being careful to disregard market impacts that result from the use of unprotected rather than protected elements.

393. Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 11 F.4th 26, 46 (2d Cir. 2021) ("Goldsmith cannot copyright Prince's face. . . . Were it otherwise, nobody else could have taken the man's picture without either seeking Goldsmith's permission or risking a suit for infringement."), *aff'd*, 598 U.S. 508 (2023).

394. *Id.*

395. *Id.* at 47.

396. *Warhol*, 598 U.S. at 525.

397. This focus on the difficulty of extraction can cut both ways. Where surgical extraction is easy, it can be unreasonable to take too much of the copyrighted work's expression. For example, in *Walt Disney Prods. v. Air Pirates*, 581 F.2d 751 (9th Cir. 1978), the Ninth Circuit explained: "[W]hen the medium involved is a comic book, a recognizable caricature is not difficult to draw, so that an alternative that involves less copying is more likely to be available than if a speech, for instance, is parodied." *Id.* at 758.

398. 17 U.S.C. § 107(4).

In general, the importance of disregarding some market impacts where they are not connected with the proper subject matter and purpose of copyright protection is well-established. Most famously, the Supreme Court explained in *Campbell v. Acuff-Rose Music, Inc.*, that "when a lethal parody, like a scathing theater review, kills demand for the original, it does not produce a harm cognizable under the Copyright Act."[399]

When a defendant's work kills demand for a plaintiff's because it represents a superior expression of the same idea (or other unprotected element), that market harm should similarly be noncognizable.[400] Although many courts have recognized this, others have lost focus on the distinction between protectable and unprotectable aspects by the time they get to the fourth factor, thus endangering the freedom to extract.

The Second Circuit articulated the importance of disregarding the market effect of competition based on unprotected elements in *Wright v. Warner Books, Inc.*,[401] in which the court determined that a biographer's use of excerpts from journal entries and letters was fair use.[402] In its analysis of the fourth factor, the court explained: "The effect on the market must be attributable to the seven instances in which the biography takes Wright's expression. Disclosure of factual content . . . is not proscribed by the copyright monopoly."[403]

A more recent Second Circuit opinion that carefully focused the fourth fair use factor on only those market effects caused by copying protectable elements is *Authors Guild v. Google, Inc.*[404] As described above, the case was a copyright challenge to Google's copying of books for purposes of generating a searchable index.[405] The search results generally included short "snippets" of the books matching the searcher's query.[406] The Second Circuit acknowledged that snippets that reveal factual information of interest to the searcher may satisfy

399. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 591–92 (1994).

400. *Cf.* Hoehling v. Universal City Studios, Inc., 618 F.2d 972, 980 (2d Cir. 1980) (rejecting both copyright and unfair competition claims based on alleged copying of historical facts and theories, explaining that "historical facts, themes, and research have been deliberately exempted from the scope of copyright protection to vindicate the overriding goal of encouraging contributions to recorded knowledge").

401. 953 F.2d 731 (2d Cir. 1991).

402. *Id.* at 733.

403. *Id.* at 739; *see also* Arica Inst., Inc. v. Palmer, 970 F.2d 1067, 1078–79 (2d Cir. 1992); Lathan v. City of Whittier Alaska, No. 3:10-CV-00070-TMB, 2011 WL 13115649, at *14 (D. Alaska Aug. 4, 2011), *aff'd*, 512 F. App'x 726 (9th Cir. 2013).

404. 804 F.3d 202, 223–24 (2d Cir. 2015).

405. *Id.* at 208.

406. *Id.* at 209.

the demand that would otherwise have been satisfied by buying the book.[407] But because this competition is based on unprotected facts, it does not count:

> [T]he type of loss of sale envisioned . . . will generally occur in relation to interests that are not protected by the copyright. A snippet's capacity to satisfy a searcher's need for access to a copyrighted book will at times be because the snippet conveys a historical fact that the searcher needs to ascertain.[408]

The Second Circuit has not always been so discerning, however. In *Warhol v. Goldsmith*, the court analyzed whether the Warhol Foundation's licensing of Warhol's depictions of the musician Prince competed with plaintiff Goldsmith's licensing market for her original photograph.[409] In holding that there was a likely market effect and that the fourth factor thus favored the plaintiff, the court wrote: "As Goldsmith succinctly states: 'both [works] are illustrations of the same famous musician with the same overlapping customer base.' Contrary to AWF's assertions, that is more than enough."[410] That should not be enough. The fact that both works are illustrations of the same famous musician means only that their unprotected factual subject matter (Prince's appearance) is the same. And the fact that they have an overlapping customer base may be only because there are many customers in the market for depictions of that subject matter. This is far from enough to establish that there was a market effect due to Warhol's borrowing of Goldsmith's expression.[411]

When the Supreme Court took up *Warhol v. Goldsmith*, it considered only the first factor and thus did not consider what type of market effect is appropriately part of the fourth factor analysis.[412] But the way the Court analyzed the first factor, focusing on whether the defendant was pursuing the same ultimate purpose as the plaintiff (versus a different, "transformative" purpose), without regard to whether it was doing so by extracting unprotected

407. *Id.* at 224.

408. *Id.*; *see also id.* at 226 (observing in the context of its factor four analysis that "[n]othing in the statutory definition of a derivative work, or of the logic that underlies it, suggests that the author of an original work enjoys an exclusive derivative right to supply information about that work of the sort communicated by Google's search functions"). For a discussion of the "protected aspects" approach to the fourth factor, see Sobel, *supra* note 47, at 55–57. Some courts cabin their consideration of competition by insisting that competition does not count against a defendant where the competition is in a "transformative market." *See, e.g.*, Apple Inc. v. Corellium, Inc., No. 21-12835, 2023 WL 3295671, at *12 (11th Cir. May 8, 2023); Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 614–15 (2d Cir. 2006).

409. 11 F.4th 26, 61–62 (2d Cir. 2021) (citation omitted), *aff'd*, 598 U.S. 508 (2023).

410. *Id.* at 49–50 (quoting Appellants' Brief, at 50, *Warhol*, 11 F.4th 26 (No. 19-2420-cv)).

411. The Second Circuit mentioned but did not decide the issue of which party has the burden of proof on this issue. *Id.* at 49 ("[W]hatever the scope of Goldsmith's initial burden, she satisfied it here.").

412. *Warhol*, 598 U.S. at 525.

elements,[413] invites the type of flawed fourth factor analysis the Second Circuit conducted.

Indeed, the Supreme Court itself has fallen prey to the temptation to consider market effects caused by copying of unprotected elements. The Court arguably committed this error in *Harper & Row, Publishers, Inc. v. Nation Enterprises*.[414] In holding that *The Nation*'s publication of an article including excerpts of President Ford's forthcoming memoir was not fair use, the majority emphasized that by "scooping" the memoir the article had undermined the market for it.[415] Evidence of this harm included the fact that *Time* canceled its own agreement to publish excerpts of the memoir.[416] The Court warned that "a fair use doctrine that permits extensive prepublication quotations from an unreleased manuscript without the copyright owner's consent poses substantial potential for damage to the marketability of first serialization rights in general."[417] The Court mentioned but did not seriously grapple with the argument, accepted by the court of appeals, that any market damage might have been due to the overlap in unprotected facts as opposed to expression.[418]

Justice Brennan's dissent in *Harper & Row* argued that the failure to carefully distinguish protected from unprotected elements "permeate[d] every aspect of the Court's fair use analysis."[419] With regard to the fourth factor, he argued:

> The Nation's publication indisputably precipitated Time's eventual cancellation. But that does not mean that The Nation's use of the 300 quoted words caused this injury to Harper & Row. Wholly apart from these quoted words, The Nation published significant information and ideas from the Ford manuscript. If it was this publication of information, and not the publication of the few quotations, that caused Time to abrogate its serialization agreement, then whatever the negative effect on

413. *Id.*; *see also* Ard, *supra* note 51 (manuscript at 43) (observing that the Supreme Court's focus on substitutability in *Warhol* appears to consider substitution based on unprotected elements); *id.* (manuscript at 54–55) (noting critique of *Warhol* in light of consensus that photographers cannot copyright faces).

414. 471 U.S. 539, 548 (1985).

415. *Id.* at 562.

416. *Id.* at 541.

417. *Id.* at 569.

418. *Id.* at 567. In another case addressing the market impact of publishing newsworthy and fact-heavy material, the trial court's analysis was similarly unclear regarding the precise basis for the competitive impact. *See* Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537, 554–55 (S.D.N.Y. 2013).

419. *Harper & Row*, 471 U.S. at 590 (Brennan, J., dissenting).

the serialization market, that effect was the product of wholly legitimate activity.[420]

By focusing on the freedom to extract, courts might avoid the mistake Justice Brennan accused the *Harper & Row* majority of making: basing the fourth factor analysis on market effects from reuse of unprotected material.

C. *Beyond the Fair Use Factors: Extractability and the Progress of Science*

So far, the adjustments to analysis of the fair use factors that I have suggested are variations on approaches that particularly careful courts, or at least individual judges or justices, have taken in the past. My final proposal for improving fair use is more novel, less factor-specific, and especially relevant to the contemporary controversies to be discussed in Part III.

Part I articulated the value and doctrinal pedigree of the freedom to extract. The fundamental claim is that safeguarding the freedom to extract, and thus preserving the idea/expression distinction itself, promotes the progress of knowledge ("science") that copyright law is supposed to foster. Part I also demonstrated that the ease of extraction can vary depending on the type of work and the form in which it is disseminated. Some works (software disseminated only in object code, works wrapped in technological protection measures, etc.) obfuscate their unprotected elements and thus make extraction difficult.[421]

The upshot of those two points combined is that the most valuable uses of existing copyrighted works do not merely extract. They make it possible for subsequent users to extract as well, and thus to benefit from the dissemination of unprotected ideas, methods, and facts.[422] On this view, a defendant who copies a photograph as evidence of an historical fact that she communicates freely to the public on Wikipedia should have a stronger fair use argument (all other things equal) than a defendant who sells access to that photograph as part of its subscription news aggregation service.[423] Similarly, a software developer who copies a program in order to extract interfaces necessary for interoperability and then publishes the source code for her own interfaces is

420. *Id.* at 602.

421. *See* Samuelson & Scotchmer, *supra* note 94, at 1661–62 (discussing the phenomenon of firms thwarting reverse engineering by designing their products to be difficult or impossible to reverse engineer and exploring potential policy responses).

422. *Cf.* Lydia Pallas Loren, *Abandoning the Orphans: An Open Access Approach to Hostage Works*, 27 BERKELEY TECH. L.J. 1431, 1458 (2012) (proposing immunity from monetary liability for actors who disseminate orphan works on open access terms and accompanied by information about the works' status).

423. *See* Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169, 174 (2d Cir. 2018) (holding that unauthorized distribution of television programming by subscription service was not fair use); Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537, 541, 554–55 (S.D.N.Y. 2013) (holding that unauthorized distribution of excerpts of news stories by subscription service was not fair use).

even more worthy of the protection of fair use (all other things equal) than a defendant who carefully guards her interfaces in secret source code.

I propose, therefore, that *extractability* should be considered as part of the fair use analysis. If a defendant's use of a copyrighted work is designed to extract unprotected elements from a copyrighted work *and* to ensure that those unprotected elements are publicly available for others to use, that should weigh even more heavily in favor of fair use than extractive use alone.[424] Extraction need not be exploitative—in the sense of taking from the intellectual commons without contributing back—if it is accompanied by extractability.

Considering extractability in this way is consistent with the Supreme Court's insistence that "we must take into account the public benefits the copying will likely produce," as part of the fair use analysis.[425] Not every public benefit is relevant to the fair use analysis (nor every public harm); fair use, as with all of copyright, should be applied to generate the specific benefit of promoting the progress of knowledge. Extractability is particularly relevant to this fundamental goal of copyright, however. By considering extractability as part of the fair use analysis, courts can ensure that extractive use promotes sustainable intellectual progress. Extractable extractive use can facilitate extraction by others and thereby further the collective project of knowledge transmission and generation.

III. EXTRACTION BY MACHINES: FREEDOM TO EXTRACT IN THE AGE OF AI

Rapidly proliferating AI technology raises the stakes for the freedom to extract and for extractability. This part explains how both extraction and extractability are relevant to contemporary copyright controversies over generative AI.

Developers of generative AI applications using models that have been trained on copyrighted works have been accused of infringing the copyrights in those works. These copyright infringement lawsuits are still in their early stages,[426] as is scholarly analysis of them. But one emerging view is that many

424. *Cf.* Edward Lee, *Technological Fair Use*, 83 S. CAL. L. REV. 797, 817 (2010) [hereinafter Lee, *Technological Fair Use*] (explaining how "[t]echnological fair uses have the potential to provide additional engines of free expression . . . given the possibility that the new speech technology in question can affect an exponential number of speech activities among the millions of people using that technology").

425. Google LLC v. Oracle Am., Inc., 593 U.S. 1, 35 (2021).

426. Generative AI copyright infringement lawsuits include: Andersen v. Stability AI Ltd., 700 F. Supp. 3d 853 (N.D. Ca. 2023); Basbanes v. Microsoft Co., 345 F.R.D. 585 (S.D.N.Y. 2024); Chabon v. Meta Platforms, No. 3:23-cv-04663 (N.D. Cal. July 22, 2024); Concord Music Grp. v. Anthropic PBC, No. 3:23-cv-01092, 2024 WL 3101098 (M.D. Tenn. June 24, 2024); Getty Images (US), Inc. v. Stability AI, Inc., 2024 WL 4262617 (D. Del. Aug. 21, 2024); Huckabee v. Meta Platforms, No.

intermediate activities in the "generative AI supply chain"[427] that involve copying[428] copyrighted works are likely to be deemed fair uses.[429] Particular

1:24CV00773, 2024 WL 3278790 (D. Del. July 1, 2024); Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC (TSH), 2024 WL 235199 (N.D. Cal. Jan. 22, 2024); J.L. v. Alphabet Inc., No. 23-cv-03440-AMO, 2024 WL 3282528 (N.D. Cal. June 6, 2024); Authors Guild v. OpenAI, Inc., 345 F.R.D. 585 (S.D.N.Y. 2024); Silverman v. OpenAI, Inc., No. 3:23-cv-03416-AMO, 2023 WL 10673219 (N.D. Cal. Sept. 27, 2023); N.Y. Times Co. v. Microsoft Corp., No. 1:23-cv-11195-SHS, 2024 WL 3691067 (S.D.N.Y. July 1, 2024); Thomas Reuters Enter. Ctr. GmbH v. ROSS Intel. Inc., 529 F. Supp. 3d 303 (D. Del. 2021); Tremblay v. OpenAI, Inc., Nos. 23-cv-03223-AMO, 23-cv-03416-AMO, 2024 WL 557720 (N.D. Cal. Feb. 12, 2024). *See generally DAIL—the Database of AI Litigation*, GEO. WASH. UNIV., https://blogs.gwu.edu/law-eti/ai-litigation-database [https://perma.cc/9XJ6-9AKZ/] (last updated Dec. 9, 2024) (database of information about litigation involving artificial intelligence).

427. *See* Katherine Lee, A. Feder Cooper & James Grimmelmann, *Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain*, J. COPYRIGHT SOC'Y U.S.A (forthcoming) (manuscript at 10–22), https://ssrn.com/abstract=4523551 [https://perma.cc/AJ2E-9Z9E (staff-uploaded archive)].

428. There is some lack of consensus about whether these activities constitute prima facie copyright infringement. *Compare, e.g.*, Lemley & Casey, *supra* note 3, at 745 (asserting that robots are "granted broad latitude" even though they copy "millions of images, videos, audio, or text-based works" which "are almost all copyrighted"), *and* A. Feder Cooper & James Grimmelmann, *The Files Are in the Computer: On Copyright, Memorization, and Generative AI*, CHI. KENT L. REV. (forthcoming), https://ssrn.com/abstract=4803118 [https://perma.cc/7HK3-UJN8 (staff-uploaded archive)] (explaining why models constitute copies of the works on which they are trained), *with* Michael D. Murray, *Generative AI Art: Copyright Infringement and Fair Use*, 26 SMU SCI. & TECH. L. REV. 259, 285 (2023) ("[C]reators of generative AI datasets . . . have not copied or stored any image files.").

429. *See, e.g.*, Pamela Samuelson, *Fair Use Defenses in Disruptive Technology Cases*, 71 UCLA L. REV. 1484, 1556–63 (2024); Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 FORDHAM L. REV. 1887, 1921 (2024); Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley & Percy Liang, *Foundation Models and Fair Use* (manuscript at 2) (Stanford L. & Econ. Olin, Working Paper No. 584, 2023), https://ssrn.com/abstract=4404340 [https://perma.cc/HXL4-BZRQ (staff-uploaded archive)]; Lemley & Casey, *supra* note 3, at 776–79; Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 619–30 (2018); Andrew W. Torrance & Bill Tomlinson, *Training Is Everything: Artificial Intelligence, Copyright, and "Fair Training,"* 128 DICK. L. REV. 223, 244–45 (2023); Jessica L. Gillotte, *Copyright Infringement in AI-Generated Artworks*, 53 U.C. DAVIS L. REV. 2655, 2680 (2020); *see also* CONG. RSCH. SERV., LSB10922, GENERATIVE ARTIFICIAL INTELLIGENCE AND COPYRIGHT LAW 3–4 (Sept. 29, 2023), https://crsreports.congress.gov/product/pdf/LSB/LSB10922 [https://perma.cc/3MUE-HPCY (staff-uploaded archive)] (discussing fair use arguments); Mehtab Khan & Alex Hanna, *The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability*, 19 OHIO ST. TECH. L.J. 171, 209–18 (2023) (discussing fair use arguments); Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training*, 74 EMORY L.J. (forthcoming 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4976393 [https://perma.cc/H6JA-ZSL6 (staff-uploaded archive)] (observing that an international equilibrium is emerging that permits unauthorized AI training on copyrighted works under some circumstances). *But see* Sobel, *supra* note 47, at 96–97 (concluding that some machine learning is unlikely to qualify as fair use); Jacob Alhadeff, Cooper Cuene & Max Del Real, *Limits of Algorithmic Fair Use*, 19 WASH J.L. TECH. & ARTS, Winter 2024, at 1 (concluding that factors tilt against fair use for text-to-image generative AI); David W. Opderbeck, *Copyright in AI Training Data: A Human-Centered Approach*, 76 OKLA. L. REV. 951, 1022–23 (2024) (arguing that AI training on copyrighted works should be based on consent, not fair use); Robert Brauneis, *Copyright and the Training of Human Authors and Generative Machines*, 47 COLUM. J.L. & ARTS (forthcoming 2025) (manuscript at 1–2), https://ssrn.com/abstract=4909592 [https://perma.cc/UPE3-RFGL (staff-uploaded archive)] (rejecting prominent arguments in favor of viewing AI training as fair use and concluding that the case for fair use is weak).

outputs that are substantially similar to works in the training data (due in part to an AI phenomenon often referred to as "memorization") may not be.[430]

Much of the reasoning that underlies these conclusions resonates with what I have said about the freedom to extract. Machine learning models appear to extract patterns from existing works. At least some of these patterns are not themselves protectable expression, but instead are facts about expression, ideas about expression, or methods of expression that allow the models to learn *how* human expression is constructed.[431] The proposals I have made for improving copyright doctrine in order to safeguard the freedom to extract are likely to help courts see the extractive value of copying related to generative AI.

In particular, the theory of extractive use developed above could help to frame the fair use analysis where artificial intelligence is used to extract unprotectable aspects from copyrighted works in order to build tools that generate new works. Such a use is not clearly transformative as that concept has most recently been understood by the Supreme Court—that is, to mean a use that serves a purpose distinct from the purpose of the original work. Generative AI can have the *same* ultimate purpose as the copyrighted work: communication of expression that can be enjoyed by readers, viewers, etc.[432] And yet, such use might legitimately be considered *extractive*, insofar as it is designed to exploit only unprotected aspects of the works it copies.

But generative AI does not only raise the stakes for extraction in terms of the innovative potential of harvesting existing works for unprotectable insights. It also raises the stakes in terms of the potential for competitive harm to

430. *E.g.*, Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295, 295–96 (2023) [hereinafter Sag, *Safety*]; Henderson et al., *supra* note 429, at 1; Nicholas Carlini, Chang Liu, Ulfar Erlingsson, Jernej Kos & Dawn Song, *The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks*, USENIX, Aug. 14–16, 2019, at 1, https://arxiv.org/abs/1802.08232 [https://perma.cc/V7FA-SXZ4]; Cooper & Grimmelmann, *supra* note 428 (manuscript at 1–2). Cooper and Grimmelmann explain that although concern about memorization often focuses on outputs that are substantially similar to works in the training corpus, the memorization itself takes place during model training. *Id.* (manuscript at 23–24).

431. *See* Lemley & Casey, *supra* note 3, at 772 (explaining that machine learning systems "generally copy works, not to get access to their creative expression . . . but to get access to the uncopyrightable parts of the work—the ideas, facts, and linguistic structure of the works"); Levendowski, *supra* note 430, at 626–27 (explaining that AI systems use fictional works to learn "abstract concepts about language or images," not to leverage the creative components of those works). *But cf.* Sobel, *supra* note 47, at 69–70 (arguing that generative AI applications extract expressive features from the corpora on which they are trained); Cooper & Grimmelmann, *supra* note 428 (manuscript at 6) (explaining that "[i]f a generative-AI model memorizes its training data, the training data is in the model"); *id.* (manuscript at 37) (explaining that "the 'patterns' learned by a model can be highly abstract, highly specific, or anywhere in between.").

432. It may also simply be unclear, when the copying involved in training an AI system is done, what the ultimate purpose(s) of that system will be. *See* Ard, *supra* note 51 (manuscript at 5–6, 9).

creators,[433] and harm to society when the bases for and implications of AI innovations are obscured from public examination. In the age of generative AI, the benefit of the freedom to extract may depend not only on machine extraction being recognized as a valid exercise of that freedom, but also on the insistence that lawful *exercise* of the freedom to extract depends in part on *facilitation* of the freedom to extract by others—that is, on extractability.

As described at the end of Part II, this means that the law should favor extractions that lay bare to the public the insights that are gleaned from preexisting works and the origins of those insights. It should look less favorably on extractions that are not disseminated at all or are disseminated in ways that obscure their unprotected elements and their sources.

Extraction paired with obfuscation is common in the world of AI, but it need not be.[434] Generative AI tools can be designed to better attribute their

433. *See* Sobel, *supra* note 47, at 77 (arguing that "[e]xpressive machine learning not only jeopardizes the market for the works on which it is trained, it also threatens to marginalize authors completely").

434. *See generally* Adrien Basdevant, Camille Fracois, Victor Storchan, Kevin Bankston, Ayah Bdeir, Brian Behlendorf, Merouane Debbah, Sayash Kapoor, Yann LeCun, Mark Surman, Helen King-Turvey, Nathan Lambert, Stefano Maffulli, Nik Marda, Gorvide Shivkumar & Justine Tunney, *Towards a Framework for Openness in Foundation Models: Proceedings from the Columbia Convening on Openness in Artificial Intelligence* (May 17, 2024), https://arxiv.org/pdf/2405.15802 [https://perma.cc/J59Q-APUC] (describing how each component of an AI system can contribute to openness and the potential benefits of AI openness); Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman, Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhang, Daniel E. Ho, Percy Liang & Arvind Narayanan, *On the Societal Impact of Open Foundation Models* (Feb 27, 2024), https://arxiv.org/abs/2403.07918 [https://perma.cc/E745-SXJ7] (describing benefits and risks of open foundation models with broadly available model weights); Andreas Liesenfeld, Alianda Lopez & Mark Dingemanse, *Opening Up ChatGPT: Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators*, INT'L CONF. ON CONVERSATIONAL USER INTERFACES (July 8, 2023), https://arxiv.org/pdf/2307.05532 [https://perma.cc/9RG7-7C8H] (documenting degrees of openness across various components of different "open-source instruction-tuned text generators"); Matt White, Ibrahim Haddad, Cailean Osborne, Xiao-Yang Lieu Yanglet, Ahmed Abdelmonsef & Sachin Matthew Varghese, *The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence* (Oct. 18, 2024), https://arxiv.org/pdf/2403.13784 [https://perma.cc/K8HV-82BC] (presenting system to rate machine learning models based on their openness); Edd Gent, *The Tech Industry Can't Agree on What Open-Source AI Means. That's a Problem.*, MIT TECH. REV. (Mar. 25, 2024), https://www.technologyreview.com/2024/03/25/1090111/tech-industry-open-source-ai-definition-problem/ [https://perma.cc/VF99-Z73L] (describing effort to define what "open-source" means for AI); Commission Regulation 2024/1689, art. 53, 2024 O.J. (L) (EU) (requiring providers of general purpose artificial intelligence to, inter alia, "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model"); Alek Tarkowsky, *AI Act Fails to Set Meaningful Dataset Transparency Standards for Open Source AI*, OPEN FUTURE (Mar. 7, 2024), https://openfuture.eu/blog/ai-act-fails-to-set-meaningful-dataset-transparency-standards-for-open-source-ai/ [https://perma.cc/J37K-FYYP] (critiquing European AI Act's approach to transparency and open-source AI). Regarding the limitations and potential unforeseen consequences of openness in AI, see David Widder, Meredith

sources.[435] Training data can be well-documented.[436] "Model cards" can describe the purpose and function of AI models.[437] Tools can be available for free public use instead of only by subscription.[438] Of course, it will not be in the competitive interest of every actor in the generative AI ecosystem to fully facilitate extraction by others. My point is only that doing so should bolster arguments against copyright liability, because extraction is fairer when it is accompanied by extractability.

CONCLUSION

This Article has documented the freedom to extract in copyright law and argued that it is essential to ensuring that copyright fulfills its constitutional purpose. I have also proposed ways in which the law could be adjusted to better safeguard the freedom to extract. The most novel adjustment that I propose is that fair use analysis should take account of both whether defendants are exercising the freedom to extract unprotected elements from copyrighted works *and* whether they are enabling others to exercise that freedom as well. This argument has important implications for artificial intelligence, which often

Whittaker & Sarah West, *Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI*, NATURE (forthcoming) (manuscript at 2–3), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807 [https://perma.cc/KM48-S8Q4 (staff-uploaded archive)].

435. *See, e.g.*, Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III & Kate Crawford, *Datasheets for Datasets*, 64 COMMS. ACM 86, 86 (2021) ("[W]e propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks."); Henderson et al., *supra* note 429, at 23; Sag, *Safety*, *supra* note 430, at 340–41 ("Those who use copyrighted works as training data for LLMs should keep detailed records of the works and from where they were obtained. . . . At a minimum, developers should keep logs and give copyright owners practical tools to determine whether their works are part of the training data."). *But cf. id.* at 340 ("LLMs that pose a significant risk of copyright infringement should not be open-sourced. If an LLM is likely to be used to generate pseudo-expression that infringes on copyrights (or other analogous rights) in a material fashion, that model should not be left unsupervised.").

436. *See* Katherine Lee, Daphne Ippolito & A. Feder Cooper, *The Devil Is in the Training Data*, *in* AI AND LAW: THE NEXT GENERATION, 5, 15 (2023), https://genlaw.github.io/explainers/training-data.html [https://perma.cc/FHZ5-8ABG] (describing a "significant push within the [machine learning] community for dataset creators to document their datasets before releasing them").

437. *See* Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji & Timnit Gebru, *Model Cards for Model Reporting*, FAT* '19: PROC. CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 220, 220–21 (2019), https://doi.org/10.1145/3287560.3287596 [https://perma.cc/U27U-ULR8].

438. *See* Lee, *Technological Fair Use*, *supra* note 424, at 848 (arguing that "it is relevant to fair use whether the technology in question is offered for free or instead at a high price to the public").

involves both extraction and obfuscation. The strongest insulation from copyright infringement should be reserved for those who extract but do not merely exploit; they enrich the public domain through extractability just as they are enriched by extraction.