BACK TO THE FUTURE: NAVIGATING THE COPYRIGHT/CONTRACT INTERFACE IN THE GENERATIVE AI ERA

Niva Elkin-Koren †

TABLE OF CONTENTS

I. INTRODUCTION

How should laws and regulations adapt to new disruptive technologies? This question poses the greatest challenge for lawyers and policymakers in the face of technological change. Meeting this challenge requires a thorough understanding of the technological transformation, including its social implications and the risks and opportunities it presents. It also involves tailoring legal rights, procedures, and institutions to address these changes.

DOI: https://doi.org/10.15779/Z38NP1WK6J

† Professor of Law, Stewart and Judy Colton Chair in Legal Theory and Innovation, Tel-Aviv University Faculty of Law, Faculty Associate, Berkman Klein Center at Harvard University; Director of the Chief Justice Meir Shamgar Center for Digital Law and Innovation, and a co-director of the Algorithmic Governance Lab at TAU Innovation Lab. Thanks to Yuval Kadari and Tamar Burstein for research assistance. I thank Assaf Hamdani, Uri Hacohen, Neil Netanel, Maayan Perel, Ohad Somech, and Molly Van Houweling for their helpful comments on earlier drafts. This Research was supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 1870/21).

This challenge has become particularly timely with the rapid rise of Generative Artificial Intelligence (GenAI). While GenAI disrupts numerous domains, including medicine, finance, transportation, and education, significant attention has been focused on its impact on the creative industries.[1] While it opens new avenues for artistic expression and collaboration, GenAI has also sparked a wave of new copyright litigation.[2] GenAI disrupts copyright law because it introduces powerful tools of generating expressive works, thus transforming how we create, share, and draw upon works created by others.[3] It raises new questions about creativity and originality, authorship and ownership, and the appropriate scope of protection and fair use for works generated with such models.[4]

The need to adapt copyright to new technology is not a new phenomenon. A few decades ago, in the early 1990s, copyright law found itself on the front lines of another technological revolution: digital technology and the internet. At the time, some believed that digital networks threatened to signal the 'death of copyright,' and, worse, of creativity as we knew it.[5]

In this context, Pamela Samuelson was among the first to accurately identify the true nature of digital transformation, highlighting the new opportunities and risks it posed, and outlining strategies for courts and policymakers on how to address them. As demonstrated in this Article, this approach may still prove useful in navigating the current disruption caused by GenAI.

---

1. *See, e.g.*, Ryan Abbott & Elizabeth Rothman, *Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence*, 75 FLA. L. REV. 1141 (2023); Oren Bracha, The Work of Copyright in the Age of Machine Production (Sept. 24, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4581738.

2. *See, e.g.*, Doe 1 v. GitHub, Inc., 672 F. Supp. 3d 837 (N.D. Cal. 2023); Authors Guild v. OpenAI Inc., No. 1:23-cv-08292, 2024 U.S. Dist. LEXIS 59322 (S.D.N.Y. Sept. 19, 2023); Andersen v. Stability AI, Ltd., 700 F. Supp. 3d 853 (N.D. Cal. 2023); Getty Images v. Stability AI, 23-CV-00135 (D. Del. 2023); N.Y. Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023).

3. *See, e.g.*, Pamela Samuelson, *Generative AI Meets Copyright*, 381 SCIENCE 158 (2023); Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley & Percy Liang, *Foundation Models and Fair Use*, STAN. L. & ECON. OLIN WORKING PAPER NO. 584 (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4404340#.

4. *See generally* Dan L. Burk, *Cheap Creativity and What It Will Do*, 57 GA. L. REV. 1669, 1673 (2023); Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295 (2023); Uri Y. Hacohen & Niva Elkin Koren, *Copyright Regenerated: Harnessing GenAI to Measure Originality and Copyright Scope*, 37 HARV. J.L. & TECH. 555 (2024); Benjamin Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45 (2017).

5. *See* John Perry Barlow, *The Economy of Ideas*, WIRED (Mar. 1, 1994), https://www.wired.com/1994/03/economy-ideas/.

Samuelson was a pioneer in identifying the rise of private ordering as the single most important implication of the digital transformation. She demonstrated how this shift has transferred the power to set legal norms from public bodies—the legislature and courts, to rights holders.[6] This shift has sparked heated debate regarding whether and under what circumstances courts should enforce contractual provisions that conflict with copyright policy.

In a series of articles, Samuelson developed a framework for addressing these issues, proposing principles to guide courts in determining when copyright should override contractual restrictions.[7] In essence, she argued that mass-market license restrictions, which undermined the core objectives of copyright policy, should be overridden.[8] This straightforward principle, guided by copyright fundamentals, offers a useful standard for navigating the copyright/contract interface. This approach is gaining increasing relevance in addressing some of the emerging copyright/contracts disputes in the new GenAI landscape.

This Article proceeds as follows: Part II describes the rise of private ordering as a dominant framework for governing creative works and explains the complex tension between copyright and contracts. It further discusses the theoretically informed, yet pragmatic approach proposed by Samuelson for resolving such conflicts. Part III demonstrates how this approach could assist in resolving emerging controversies involving the use of copyrightable materials for training GenAI systems. Finally, Part IV concludes.

6. *See, e.g.*, Pamela Samuelson, *Copyright and Freedom of Expression in Historical Perspective*, 10 J. INTELL. PROP. L. 319 (2003); Pamela Samuelson, *Intellectual Property and the Digital Economy: Why the Anti-circumvention Rules Need to Be Revised*, 14 BERKELEY TECH. L.J. 519 (1999); Pamela Samuelson, *Intellectual Property Rights and the Global Information Economy*, 39 COMM. ACM. 23 (Jan. 1996); Pamela Samuelson, *The Copyright Grab*, 4.01 WIRED 134 (Jan. 1996).

7. *See* Pamela Samuelson & Suzanne Scotchmer, *The Law and Economics of Reverse Engineering*, 111 YALE L.J. 1575, 1661 (2001); Pamela Samuelson, Jon A. Baumgarten, Michael W. Carroll, Julie E. Cohen, Troy Dow, Brian Fitzgerald, Laura Gasaway, Daniel Gervais, Terry Ilardi, Jessica Litman, Lydia Pallas Loren, Glynn Lunney, Tyler Ochoa, R. Anthony Reese, Jule Sigall, Kate Spellman, Christopher Sprigman, Michael Traynor, Tara Wheatland & Jeremy Williams, *The Copyright Principles Project: Directions for Reform*, 25 BERKELEY TECH. L.J. 1175, 1235–38 (2010); Pamela Samuelson, *Possible Futures of Fair Use Symposium*, 90 WASH. L. REV. 815 (2015); Pamela Samuelson, *Freedom to Tinker*, 17 THEORETICAL INQUIRES L. 563 (2016).

8. Samuelson, *Possible Futures of Fair Use Symposium*, *supra* note 7, at 859–60.

II. THE DIGITAL TRANSFORMATION AND THE PRIVATIZATION OF COPYRIGHT

A. THE RISE OF COPYRIGHT PRIVATE ORDERING

The advent of digital technology caused significant disruption in copyright law, a legal regime that originally developed alongside the introduction of the printing press.[9] While there was a consensus that the law must adapt to confront these fundamental challenges, ongoing controversies persisted regarding the nature of these challenges and the necessary adaptations required in copyright law.[10]

Samuelson was a pioneer in identifying the rise of private ordering as the single most important implication of the digital transformation.[11] As the internet grew more popular, scholars and policymakers primarily focused on the ease of copying and costless dissemination enabled by digital networks, advocating for stronger copyright protection and further expansion of rights granted to copyright owners.[12] Yet, digital technology has also introduced new methods of governing the use of copyrighted materials.[13] Given the ease of digital coping, rights holders increasingly turn to self-help measures, utilizing technological protection measures and boilerplate contracts to protect their rights and interests.[14]

Technological protection measures, such as encryption and password-protected paywalls, were soon backed by the anti-circumvention rules

9. *See generally* RAY PATTERSON, COPYRIGHT IN HISTORICAL PERSPECTIVE (1968); MARK ROSE, AUTHORS AND OWNERS: THE INVENTION OF COPYRIGHT (1993).

10. One such issue of controversy during the 1980s was whether to protect computer programs. The United States 1978 report of the National Commission of New Technological Uses of Copyrighted Works (CONTU) argued that while programmers need to invest significant effort to create economically valuable software, the ease of digital copying threatens to undermine the incentive to make that investment. Consequently, it was argued that software should be protected by copyright as a "literary work." Samuelson, however, argued that computer programs are utilitarian works, and affording them copyright protection as literary works might be a misfit. *See* Pamela Samuelson, *CONTU Revisited: The Case Against Copyright Protection for Computer Programs in Machine-Readable Form*, 1984 DUKE L.J. 663 (1984); 93d Cong., NATIONAL COMMISSION ON NEW TECHNOLOGICAL USES OF COPYRIGHTED WORKS (1974) [hereinafter CONTU].

11. *See* Samuelson, *Possible Futures of Fair Use, supra* note 7, at 859 ("Two techniques commonly used to impede fair uses of copyrighted works are the adoption of mass-market license agreements containing restrictive terms and the implementation of technical protection measures (TPMs) that prevent access to or copying of digital works.").

12. *See id.*; *see, e.g.*, CONTU *supra* note 10.

13. *See* Samuelson, *Possible Futures of Fair Use*, *supra* note 7, at 859.

14. *See, e.g.*, Viktor Mayer-Schonberger, *Beyond Copyright: Managing Information Rights with DRM*, 84 DENV. U. L. REV. 181 (2006).

established under the Digital Millennium Copyright Act of 1998 (DMCA).[15] These rules prohibit ducking the technological protection measures used by copyright owners to control access to their works[16] and impose civil and sometimes even criminal liability for circumventing these locks.[17] Mounting criticism has emerged during the 2000s, raising concerns regarding the expansion of copyright and enclosure of the public domain under this legal regime.[18] Technological protection measures, initially forecasted to limit access to knowledge, did not prove useful for protecting copies of copyrighted works as anticipated. Nowadays, however, access to copyrighted materials is predominantly facilitated through digital platforms and relies heavily on streaming, cloud computing, and mobile access.[19] This has strengthened the scope and effectiveness of technological measures, making algorithmic management of access even more robust. In this model, users do not acquire any physical copies of works. Instead, copyrighted materials (e.g., software, music, books) are stored remotely, and users' rights and obligations are determined by technological affordances and contractual restrictions imposed by digital platforms.

A second method applied by rights holders to restrict some uses of copyrighted materials is contracts. The availability of expressive materials online enabled rights holders to directly contract with the end users of their respective works at low cost, subjecting any access to materials to contractual provisions in the Terms of Service (ToS) of the hosting facility (e.g., cloud service, digital platform). Such boilerplate contracts often stipulate that users who access or download content agree to the contractual terms simply by entering the website.

---

15. *See generally* David Nimmer, *A Riff on Fair Use in the Digital Millennium Copyright Act*, 148 U. PENN. L. REV. 673 (2000). Under 17 U.S.C. § 1201, it is a violation of the DMCA to circumvent or traffic in products meant to circumvent an access control measure used to protect a copyrighted work.

16. Pamela Samuelson, *Intellectual Property and the Digital Economy: Why the Anti-Circumvention Regulations Need to be Revised* (1999); JONATHAN L. ZITTRAIN, TECHNOLOGICAL COMPLEMENTS TO COPYRIGHT (2005).

17. 17 U.S.C. § 1203 sets civil remedies for a violation of the anti-circumvention provisions including temporary or permanent injunctions and either actual or statutory damages. Criminal liability may apply in case of willful violation of the anti-circumvention provisions for commercial advantage or private gain under 17 U.S.C. § 1204.

18. *See* Pamela Samuelson, *The Copyright Grab*, WIRED (Jan. 1, 1996); JESSICA D. LITMAN, DIGITAL COPYRIGHT (2000); JAMES BOYLE, THE PUBLIC DOMAIN: ENCLOSING THE COMMONS OF THE MIND (2008); LAWRENCE LESSIG, THE FUTURE OF IDEAS: THE FATE OF THE COMMONS IN A CONNECTED WORLD (2001).

19. Niva Elkin-Koren, *The New Frontiers of User Rights*, 32 AM. U. INT'L L. REV. 1, 2–13 (2016).

Boilerplate contracts and license agreements typically impose additional restrictions on the use of copyrighted materials, sometimes conflicting with copyright norms.[20] For instance, contractual terms may prohibit the adaptation, repairs, or modifications of software,[21] limit the right to engage in reverse engineering (which might be permissible under fair use),[22] restrict consumers' rights to resell or give away purchased copies of copyrighted works such as digital books (often protected under first-sale doctrine), or constrain the use of facts and raw data, which are otherwise not protected by copyright law.[23]

Samuelson raised concerns regarding the rise of private ordering as a dominant strategy employed by rights holders for governing creative works. Drawing upon the history of the Stationers' copyright system,[24] she warned that leaving the exploitation of informational works solely at the discretion of rights holders may undermine public policy.[25]

Are contractual terms that may conflict with copyright law legally enforceable, and should they be as a matter of legal policy? The following sections aim to explore these questions.

B. PREEMPTION DOCTRINE AND ITS LIMITS

Copyright preemption doctrine is a legal framework for scrutinizing state law claims, including breach of contract, which may conflict with copyright norms under the 1976 Copyright Act.[26] This doctrine may effectively limit the freedom of copyright owners to contract around copyright law in order to

20. *Id.* at 10.

21. Universal Instruments Corp. v. Micro Sys. Eng'g, Inc., 924 F.3d 32, 48–49 (2d Cir. 2019).

22. Sega Enterp. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1522–24 (9th Cir. 1992); Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596 (9th Cir. 2000).

23. *See* Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 346 (1991). *See* discussion *infra* notes 119–121 and accompanying text. Contractual restrictions may apply to various aspects of creative works, including both copyright-protected elements (i.e., original expressions) and those not covered by copyright (e.g., ideas). Although unprotected, these aspects are often inseparably linked with the copyrighted elements. This enables rights holders to impose broad restrictions in their license agreements, extending limitations on access and use to the unprotected aspects of the licensed works.

24. RAY PATTERSON, COPYRIGHT IN HISTORICAL PERSPECTIVE (1968); *see also* MARK ROSE, AUTHORS AND OWNERS: THE INVENTION OF COPYRIGHT (1993); JOHN FEATHER, A HISTORY OF BRITISH PUBLISHING (1988).

25. Pamela Samuelson, *Copyright, Commodification and Censorship: Past as Prologue—but to What Future?*, THE COMMODIFICATION OF INFORMATION 63 (Niva Elkin-Koren & Neil W. Nentanel eds., 2002).

26. 17 U.S.C. § 301(a).

extend any of the exclusive rights granted to them by law.[27]Accordingly, in cases where contractual provisions effectively override copyright law, this doctrine may transform the legal dispute between rights holders and potential users of copyrighted materials, from a contractual cause of action subject to state law into a federal copyright dispute.

Section 301(a) states that copyright protection arises exclusively under federal law, preempting state law claims involving rights that are "equivalent to any of the exclusive rights within the general scope of copyright" and that "come within the subject matter of copyright."[28] Accordingly, to determine whether a specific contractual claim is preempted under the law, courts apply a two-part analysis.[29] First, the court examines whether the work affected by the plaintiff's exercise of a state-created right (here, the breach of contract claim) falls within the subject matter of copyright, as specified by sections 102 and 103.[30] If the subject matter of the state law claim falls within copyright subject matter, then the court must determine whether the rights asserted under the state law claim are equivalent to any of the exclusive rights listed under section 106 of the Copyright Act.[31]

To satisfy the first prong of the analysis—the "subject matter requirement"—the underlying work must be "a 'literary work,' a 'musical work,' a 'sound recording,' or any other category of 'work of authorship' within the 'subject matter of copyright.'"[32] Importantly, the scope of copyright for the purpose of preemption is broader than the scope of available copyright protection. Hence, a plaintiff may not claim a contractual right in a *type* of work covered by section 102 and 103 of the Copyright Act, even if, for some reason, that work is not protected by copyright (i.e., because it fell into the public domain or because it lacks sufficient originality).[33] Such a broad interpretation procures the statutory distinction between copyrighted works and works that ought to remain in the public domain, limiting the ability of the parties to contract around it.

---

27. Niva Elkin-Koren, *Copyright in the Digital Ecosystem: A User Rights Approach*, in COPYRIGHT LAW IN AN AGE OF LIMITATIONS AND EXCEPTIONS 132 (Ruth Okediji ed., 2017).

28. 17 U.S.C. § 301(a) states: "[A]ll legal or equitable rights that are equivalent to any of the exclusive rights within the general scope of copyright as specified by section 106 in works of authorship that are fixed in a tangible medium of expression and come within the subject matter of copyright as specified by sections 102 and 103 . . . are governed exclusively by Title 17."

29. MLGenius Holdings LLC v. Google LLC, No. 20-3113, 2022 U.S. App. LEXIS 6206, at *4 (2d Cir. Mar. 10, 2022), *cert. denied*, 143 S. Ct. 2658 (2023).

30. *See* Jackson v. Roberts, 972 F.3d 25, 42 (2d Cir. 2020).

31. *Genius*, 2022 U.S. App. LEXIS, at *4.

32. *Jackson*, 972 F.3d at 42–43.

33. Forest Park v. Universal TV Network, Inc., 683 F.3d 424, 429–30 (2d Cir. 2012).

To meet the second prong of the preemption test—the "equivalence" requirement—the defendant must show that the right the plaintiff asserts in the work (which meets the first prong of the test) is "equivalent to any of the exclusive rights within the general scope of copyright as specified by section 106."[34] For preemption to apply, "the state law claim must involve acts of reproduction, adaptation, performance, distribution or display."[35] Nevertheless, a claim is not preempted if it "include[s] any extra elements that make it qualitatively different from a copyright infringement claim."[36] To determine whether such an extra element exists, courts evaluate "what the plaintiff seeks to protect, the theories in which the matter is thought to be protected and the rights sought to be enforced."[37] This requires "a holistic evaluation of the nature of the rights sought to be enforced, and a determination whether the state law action is qualitatively different from a copyright infringement claim."[38]

As further discussed below, the enforceability of contractual provisions that may conflict with copyright has been highly contested. Despite mounting critique by academics, studies show that courts have routinely approved such contractual override of copyright norms in numerous cases.[39] In fact, over the past decades, courts have rejected the preemption of contractual restrictions, practically upholding any contractual provision that applies to those parties who agreed to accept it.[40]

Recently, however, the Second Circuit reaffirmed preemption claims concerning contracts in two significant cases. In *Universal Instruments Corp. v. Micro Systems Engineering Inc.*, the court addressed a contractual dispute between a software provider and a licensee. The plaintiff licensed its software for the first phase of the defendant's project, while subsequent phases were eventually assigned to its competitor. The plaintiff argued that adapting the software during the implementation of the second phase was contrary to the license agreement and, therefore, constituted a breach of contract.[41] The Second

34. *Jackson*, 972 F.3d at 43.

35. *Genius*, 2022 U.S. App. LEXIS 6206, at *6.

36. *Id.*

37. *Id.* at *7.

38. *Jackson*, 972 F.3d at 44 n.17.

39. *See* Guy A. Rub, *Copyright Survives: Rethinking the Copyright-Contract Conflict*, 103 VA. L. Rev. 1141 (2017) (surveying 279 cases and proposing the "no-preemption" approach is the prevalent interpretation for copyright preemption in the United States).

40. *Id.*

41. Universal Instruments Corp. v. Micro Sys. Eng'g, Inc., 924 F.3d 32, 49 (2d Cir. 2019) (holding that a contract claim was preempted and explaining that the parties' "contractual privity does nothing to change the fact that vindication of an exclusive right under the

Circuit held that the breach-of-contract claim, which involved adaptation covered by copyright law, was preempted by the Copyright Act.

In another recent decision by the Second Circuit, *ML Genius Holdings LLC v. Google LLC*, the court reaffirmed copyright preemption of another type of breach-of-contract claim: a website's ToS. The plaintiff, Genius, a publisher of lyric transcriptions of songs, argued that Google had copied the lyrics and displayed them in its search results, in violation of its ToS. The appellate court held the contract claim was preempted because the plaintiff effectively sought to protect a right that is coextensive with the exclusive right to reproduce and make derivative works of the lyrics.[42] The Supreme Court denied the plaintiff's petition for a writ of certiorari.[43]

The scope of copyright preemption doctrine involving breach-of-contract claims has remained vague and ambiguous, largely because copyright and contracts are often complimentary. Copyright law defines initial entitlements, while contracts govern their transfer. The distinction between a property license, which derives its binding force from the grant of exclusive rights by copyright, and a contract, where the binding force stems from the parties' consent, is not always easy to discern. As further discussed below, that is especially true in the context of mass-market licenses. Therefore, when boilerplate contracts are involved, it is difficult to determine whether an extra element exists that would trump the preemption doctrine.

C. SHOULD COPYRIGHT OVERRIDE CONTRACTUAL RESTRICTIONS?

The enforceability of contractual provisions that override copyright law has been a subject of long-standing debate among legal scholars.[44] Advocates of private ordering argue that contracts are essential for facilitating market transactions, enabling rights holders to extract remuneration from the commercial exploitation of their works.[45]The copyright regime is designed to generate incentives to authors by facilitating market transactions, assuming that markets could generally allocate access to creative works in the most efficient way and signaling when creative expressions are socially valuable.

Copyright Act" asserted through a breach-of-contract claim "is preempted by the Copyright Act").

42. *See Genius*, 2022 U.S. App. LEXIS 6206, at *3.

43. *See id.*

44. *See* Amit Elazari Bar On, *Unconscionability 2.0 and the IP Boilerplate*, 34 BERKELEY TECH. L.J. 567, 595 (2019).

45. Licenses, terms of use, terms of service of digital platforms, or end users license agreements, often simply authorize specific uses in copyrightable materials. These agreements provide an important legal mechanism for exercising copyright, enabling rights holders to generate remunerations from their works.

Therefore, ensuring efficient contract formation between rights holders and potential licensees is considered a key to a thriving market in copyrighted materials. Arguably, it follows that, as a general matter, contracts should be enforceable, even when potentially expanding the rights granted to rights holders under copyright law.[46]

Advocates further argue that contracts can never contradict copyright law. This is due to a division of labor between two legal regimes: copyright laws are responsible for defining initial entitlements, while contracts govern their transfer. Copyright creates rights *in rem*, while contracts only apply to their immediate contracting parties.

This approach was established by the Seventh Circuit in *ProCD v. Zeindeberg,* where the court held that a breach-of-contract claim was not preempted because a "copyright is a right against the world," while "[c]ontracts, by contrast, generally affect only their parties" and therefore "do not create 'exclusive rights.'"[47] In *ProCD*, the plaintiff sought to protect uncopyrightable digitized telephone listings using a shrink-wrap license. The appellate court held that such contracts only affect the parties involved and cannot establish rights *in rem* equivalent to copyright. Under this line of reasoning, a breach-of-contract claim would stand unless the respondent can prove that no valid contract was formed. Issues of contract formation may arise, for instance, in "browsewrap" agreements, where questions of awareness to the contractual restrictions and explicit consent might need further elaboration.[48]

Samuelson has raised concerns over the use of contractual restrictions that contradict the norms set by copyright law, thus undermining its goals. She has warned against the potential chilling effect of enforcing such restrictive duties on the fundamental goals of copyright law, which aim to promote progress.[49]She has also expressed worries that contractual terms, providing

46. *See* Guy A. Rub, *Copyright and Copying Rights*, 98 N.Y.U. L. REV. 342 (2023).

47. ProCD, Inc. v. Zeidenberg, 86 F.3d 1447, 1454 (7th Cir. 1996); *see also* Bowers v. Baystate Techs., Inc., 320 F.3d 1317, 1325 (Fed. Cir. 2003), *cert. denied*, 539 U.S. 928 (2003); Lipscher v. LRP Publ'ns, Inc., 266 F.3d 1305, 1318 (11th Cir. 2001) (finding that "claims involving two-party contracts are not preempted because contracts do not create exclusive rights, but rather affect only their parties," and holding that a contract claim was not preempted).

48. Nguyen v. Barnes & Noble Inc., 763 F.3d 1171, 1176 (9th Cir. 2014) ("The defining feature of browsewrap agreements is that the user can continue to use the website or its services without visiting the page hosting the browsewrap agreement or even knowing that such a webpage exists.").

49. *See* Samuelson, *Copyright and Freedom of Expression in Historical Perspective, supra* note 7, at 335–36; *see* also Niva Elkin Koren, *Can Formalities Save the Public Domain? Reconsidering Formalities for the 2010s*, 28 BERKELEY TECH. L.J. 1537, 1537–38 (2013); Niva Elkin-Koren, *A*

broader protection to rights holders beyond those secured by copyright law, may compromise socially beneficial practices. For instance, when rights holders seek to restrict the fair use of creative works, such as tinkering,[50] testing, criticism, research, and learning,[51]they risk undermining the goals of copyright law, which aims to promote these uses.[52]

At the same time, however, copyrights and contracts are complementary. Copyright grants authors exclusive rights in their creative works, providing incentives to create by making these works excludable. This allows right holders to earn compensations through licensing agreements and contracts. While copyright law defines these entitlements, contract law enables rights holders to leverage them effectively to generate income. Therefore, Samuelson concluded that a more promising approach is to avoid a categorical approach that would render unenforceable any contractual term that contradicts copyright.[53] Instead, she suggested that the articulation of standards for the enforceability of such restrictive terms would evolve through court decisions.[54]

In a series of articles, Samuelson outlined several key principles to guide courts in determining under what circumstances copyright should override contractual restrictions. Essentially, she argued that restrictions in mass-market licenses that undermine the core objectives of copyright policy should be overridden. The proposed principles address three aspects: the type of contract, the identification of potential conflicts between different norms, and the rationale for overriding contractual terms based on copyright policy.

First, Samuelson drew a key distinction between negotiated contracts, where courts are unlikely to override contractual restrictions that limit fair use, and mass-market license agreements (boilerplate contracts), where courts might be more willing to override license terms.[55] The nature of the contract

---

Public-Regarding Approach to Contracting Copyrights, EXPANDING THE BOUNDARIES OF INTELLECTUAL PROPERTY: INNOVATION POLICY FOR THE KNOWLEDGE SOCIETY, 191, 192 (Rochelle C. Dreyfuss, Diane L. Zimmerman & Harry First 2001); Margaret J. Radin, *Regime Change in Intellectual Property: Superseding the Law of the State with the "Law" of the Firm*, 1 U. OTTAWA L. & TECH J. 173, 178 (2004).

50. Samuelson, *Freedom to Tinker*, *supra* note 7, at 564; Maayan Perel & Niva Elkin Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69 FLA. L. REV. 181 (2017).

51. *See* Samuelson, *Copyright, Commodification and Censorship*, *supra* note 25; *see also* Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743 (2021).

52. *See* Samuelson, *Possible Futures of Fair Use Symposium*, *supra* note 7, at 859.

53. *Id.* at 859 ("articulation of standards for determining under what circumstances fair use should override license or technical restrictions" might evolve).

54. *Id.*

55. Samuelson, *Possible Futures of Fair Use Symposium*, *supra* note 7, at 859–60 (arguing that "[i]t seems unlikely that courts would accept that fair use should either always or never override

is significant both as a matter of contract law and as a matter of copyright policy. From the perspective of contract law, holding contractual restrictions binding assumes they reflect an efficient bargain by consenting parties. However, contracts governing the use of copyrighted materials, such as browsewrap licenses (e.g., website ToS) and shrinkwrap license (e.g., End User License Agreements (EULAs)), have been enforced by courts based on very minimal evidence of consent.[56] As eloquently described by Margaret Jane Radin: "The idea of voluntary willingness first decayed into consent, then into assent, then into a mere possibility or opportunity for assent, then merely fictional assent, then to mere efficient rearrangement of entitlements without any consent or assent."[57] In the absence of meaningful consent, there is no guarantee that the contract indeed reflects an efficient allocation of rights.

The boilerplate nature of the contract also matters for copyright policy. While negotiated contractual obligations and their corresponding rights apply only to the contracting parties, mass-market contracts widely expand the contract's scope. In fact, in digital distribution, the contract and the copyrighted work often converge, making the breach of contractual restrictions universally applicable. Smart contracts take this convergence a step further, by enabling automatic enforcement. In the absence of meaningful privity, boilerplate contracts become universally enforceable. They apply to each access to the work, thus creating a *de facto* right against the world.[58] Consequently, such contracts have the same impact as property rules. This blurs the distinction between copyright (property) and contracts. Given the scope and scale of such mass-market licenses, judicial scrutiny is essential for sustaining the balance of interests that copyright seeks to achieve.[59]

A second challenge in deciding whether a contractual obligation should be overridden by copyright is identifying which contractual restrictions potentially trigger copyright policy. This requires moving beyond a technical application of the preemption doctrine's two-part test, to analyze the substance of the

---

contractual restrictions" and that "[t]he most promising approach is one that would override mass-market license restrictions that interfere with copyright policy purposes").

56. Mark A. Lemley, *Beyond Preemption: The Law and Policy of Intellectual Property Licensing*, 87 CALIF. L. REV. 111 (1999). For instance, shrinkwrap licenses were enforced even when the licensee became aware of the terms only after the computer program was purchased (*ProCD Inc.*, 86 F. 3d 1447). Similarly, browsewrap licenses were held enforceable even where the license provisions were simply posted online stating that the mere use of the product or website constituted acceptance of the terms of the license.

57. Margaret J. Radin, *Boilerplate Today: The Rise of Modularity and the Waning of Consent*, 104 MICH. L. REV. 1223, 1231 (2006).

58. *See* Niva Elkin-Koren, *Copyright Policy and the Limits of Freedom of Contract*, 12 BERKLEY TECH. L.J. 93 (1997).

59. *See* Samuelson, *The Copyright Principles Project*, *supra* note 8, at 1237.

breach of contract claim and determine whether it is ultimately governed by copyright policy. In 2010, Samuelson collaborated with a group of copyright experts to propose amendments to copyright law, aiming to adjust it to the digital transformation.[60] The team listed some factors that courts may consider in determining whether a breach of contract claim should be preempted by federal copyright law.[61] In essence, "contractual provisions that forbid undertaking activities that copyright law would otherwise permit or that require action, such as giving attribution, that copyright law otherwise does not expressly require, should be subject to implied preemption analysis in appropriate cases."[62]

Finally, in deciding whether copyright should override contractual restrictions, it is essential to assess whether enforcing these restrictions would conflict with copyright policy. Samuelson outlines three concerns that offer a useful framework for examining whether specific restrictions undermine copyright policy. One concern is chilling free speech when contracts restrict criticism, research, and learning. Another concern relates to stifling innovation, interoperability, and competition, when contractual provisions seek to limit tinkering, reverse engineering, or systems testing.[63] In this context, Samuelson proposed considering factors such as the nature of the market to which these restrictive provisions apply, specifically whether it is a market with strong

60. The Copyright Principles Project (CPP) aimed to propose some improvements to adjust copyright law to the challenges raised by the digital transformation. *Id.* at 1235–38.

61. Among the considerations recommended by the team are: (1) the extent to which the contractual provision at issue alters the scope of protection copyright would otherwise provide; (2) whether the contractual provision accompanies a work that is published or otherwise publicly distributed; (3) whether the contractual provision is individually negotiated or part of a uniform, mass-market license; (4) whether the idea or information that is the subject of contractual protection is otherwise readily available from other sources without similar contractual restrictions; (5) whether enforcing the contract would establish legal control over ideas or information that copyright leaves unprotected in ways that would unreasonably inhibit future authorship or create undue monopolization; (6) whether the contract would stifle the dissemination of new creative works, such as works that criticize or comment on existing works; (7) the copyright owner's purpose in including the challenged provision in the contract; (8) whether failure to enforce the contractual provision frustrate efficient, practical enforcement of the copyright rights; and (9) whether the contract would restrict access to works that are no longer protected by copyright. *Id.*

62. *Id.*

63. For instance, Samuelson conceptualizes the freedom to tinker, which is essential for innovation, demonstrating how contract could place substantial constraints on user rights to tinker with and modify computer programs and digital works. *See* Samuelson, *Freedom to Tinker*, *supra* note 7.

network effects.[64] Finally, contractual provisions may shrink the public domain when contracts aim to extend protection beyond what is granted by copyright law—to uncopyrightable ideas, functionalities, and data. These three considerations—free speech, innovation, and the public domain—could guide courts in identifying the circumstances where contractual restrictions should be overridden.

This straightforward standard proposed by Samuelson for navigating the copyright/contract interface may prove useful in addressing emerging copyright disputes in the GenAI era, as further discussed in Part III.

III. CONTRACT/COPYRIGHT INTERFACE & GENAI

A. CONTRACTUAL BARRIERS TO TRAINING GENAI SYSTEMS

GenAI models are rapidly expanding in popularity and reach. GenAI may bring about great benefits to society, potentially leading to important breakthroughs in various domains such as medicine, transportation, and governance.[65] These systems piggyback on the impressive capability of foundation models, such as OpenAI's Generative Pre-trained Transformer (GPT) or Google's Bidirectional Encoder Representations from Transformers (BERT), to extrapolate patterns and structures from granular data.[66]Foundation models are large-scale neural networks pre-trained on vast amounts of unlabeled data, using self-supervised learning on surrogate tasks. These general-purpose models learn generalizable and adaptable data representations that can be applied to new downstream tasks. Adapting to new tasks may involve techniques such as fine-tuning, namely training the foundation model on a smaller, task-specific dataset.[67] ChatGPT, for instance,

64. *See* Samuelson & Scotchmer, *The Law and Economics of Reverse Engineering*, *supra* note 7, at 1661 (arguing that there is no "intrinsic reason" to allow contracts to circumvent copyright "especially in markets with strong network effects.").

65. *See e.g*, Bill Gates, *The Age of AI Has Begun*, GATESNOTES (Mar. 21, 2023), https://www.gatesnotes.com/The-Age-of-AI-Has-Begun; *From Code to Cure, How Generative AI Can Reshape the Health Frontier*, DELOITTE (2023), https://www2.deloitte.com/content/dam/Deloitte/us/Documents/life-sciences-health-care/us-from-code-to-cure-1.pdf.

66. *See generally* Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley & Percy Liang, *Foundation Models and Fair Use*, STAN. L. & ECON. OLIN WORKING PAPER, Paper No. 584 (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4404340# [https://perma.cc/5RTM-NVBT; Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora & Sydney Von Arx, *On the Opportunities and Risks of Foundation Models*, ARXIV, 5–15 (2021), https://arxiv.org/abs/2108.07258.

67. Bommasani et al., *supra* note 67 at 3 ("A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks . . . .").

is built on OpenAI's foundation models GPT3.5 and GPT4 to enable bot-human interaction. ChatGPT could be fine-tuned further for more nuanced natural language processing tasks, such as language translation, classification, and text summarization.

Foundation models were initially aimed at learning about data without supervision[68]but were increasingly deployed for generative applications.[69] Downstream diffusion applications, such as Midjourney and Stable Diffusion, which have been the focus of recent copyright disputes, involve text-to-image tools for creating and editing visual works. These applications enable users to generate original expressive outputs by prompting the model with a written description of the desired output.[70] Like the foundation models on which they are based, diffusion applications do not rely on formal instructions to generate original outputs. Instead, these models' learning is extrapolated from their preexisting training examples. For example, Copilot and Midjourney are trained on giant corpora of prewritten code and images from the GitHub open-source code repository and the LAION 5B database, respectively.[71] By allowing users to generate code and images in response to "prompts", GenAI tools arguably augment human creativity and democratize the creative process by enabling users to draw upon synthesized preexisting works.[72]

The emergence of foundation models and GenAI was made possible by the availability of vast amounts of data and the advancements in data storage and processing capabilities.[73]

---

68. Sam Bond-Taylor, Adam Leach, Yang Long & Chris G. Willcocks, *Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models*, 44 IEEE TRANSCON. PATTERN ANALYSIS & MACH. INTELL. 7327, 7327 (2022).

69. *Id.*

70. *See* Bracha, *supra* note 2, at 10 ("The main purpose of GenAI, by contrast, is to generate new informational goods . . . ").

71. Romain Beaumont, *Laion-5B: A New Era of Open Large-Scale Multi-Modal Datasets*, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b/ (explaining the Laion-5B project); Doe 1 v. GitHub, Inc., 672 F. Supp. 3d 837, 846–48 (N.D. Cal. 2023); Andersen v. Stability AI, Ltd., 700 F. Supp. 3d 853 (N.D. Cal. 2023).

72. *Cf.* Rachel Metz, *AI Won an Art Contest, and Artists Are Furious*, CNN BUS. (Sept. 3, 2022), https://edition.cnn.com/2022/09/03/tech/ai-art-fair-winner-controversy/index.html (reporting that a game designer with no artistic training won a fine art competition using GenAI).

73. For example, Google's T5 and Facebook's LLaMA datasets were trained on content from over 15 million websites. Similarly, OpenAI's GPT-3 features a model size of 175 billion parameters and was trained with 45 terabytes of data. *See* Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu & Lichao Sun, *A*

Like humans, GenAI models learn from examples. They use preexisting materials to learn patterns of language, extract common styles, generalize principles from new materials, and apply these features in reconstructing a new output. They generalize expressive patterns and apply those learnings to perform different tasks, such as autocompleting sentences or generating visual outputs in response to a textual prompt. Yet, unlike human learning, which occurs within the confines of the human mind, GenAI learning involves digital reproduction. Thus, the training of GenAI models requires access and copying of massive amounts of data, often scraped from the internet.

The scraping of materials while training AI systems has sparked numerous class actions lawsuits alleging copyright infringement. These lawsuits claim that the models infringe on copyright by copying protected materials without authorization from the rights holders.[74] Scraping involves "'extract[ion of] data from a website' and copying it into a structured format, allowing for data manipulation or analysis."[75] The process is not intended to store or distribute the original content but to temporarily copy it to extract data and learn patterns. Therefore, scraping for the purpose of training AI models has invoked a heated debate among scholars and policy makers.[76] Some believe that the use of copyrighted materials in the course of training GenAI models generally amount to fair use.[77]Others are more skeptical, claiming that the use of copyrighted materials during training does not amount to what courts have traditionally labeled non-infringing non-expressive use.[78]Some scholars argue that since such intermediary scraping, intended to simply extract data and learn patterns that are otherwise not protected by copyright law, should not even be considered an infringement of the exclusive right to copy.[79]

Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT, ARXIV (May 2023), https://arxiv.org/pdf/2302.09419.pdf.

74. *See* Samuelson, *Generative AI Meets Copyright*, *supra* note 3; *see, e.g.*, Andersen v. Stability AI, Ltd., No. 23-cv-00201-WHO, 2024 U.S. Dist. LEXIS 143204, at *1, *31 (N.D. Cal. Aug. 12, 2024); Tremblay v. OpenAI, Inc., Nos. 23-cv-03223-AMO; 23-cv-03416-AMO, 2024 U.S. Dist. LEXIS 24618, 1, 11–12 (N.D. Cal. Feb. 12, 2024).

75. *See* hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 991 n. 3 (9th Cir. 2019).

76. *See generally* Lemley & Casey, *Fair Learning*, *supra* note 51; Benjamin L. W. Sobel, *A New Common Law of Web Scraping*, 25 LEWIS & CLARK L. REV. 147 (2021).

77. *See, e.g.*, Lemley & Casey, *Fair Learning*, *supra* note 51, at 767; OPINION: USES OF COPYRIGHTED MATERIALS FOR MACHINE FOR MACHINE LEARNING, Israeli Ministry of Justice (2022), https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf (the use of copyrighted content to train a machine-learning tool is likely to be permissible as "fair use" under copyright law).

78. *See* Sobel, *supra* note 5, at 48.

79. *See, e.g.*, Bracha, *supra* note 2.

Uncertainty regarding copyright claims involving scraping has led rights holders to undertake self-help measures to protect their content against it. Many stakeholders enforce contractual obligations prohibiting scraping and supplement these with technological measures to detect or prevent such scraping. For example, the ToS of digital platforms often prohibit automated data collection of publicly available data, such as users' posts or personal data hosted by the platform.[80]Similarly copyright owners use their EULA to restrict the use of their works in the course of training GenAI models.[81]

This part describes these contractual practices and examines whether such contractual obligations should be enforceable.

B. RESTRICTIONS ON SCRAPING IN DIGITAL PLATFORMS

Digital platforms often prohibit the scraping of data hosted by the platform in their ToS. Data, often referred to as the 'new oil',[82] or 'gold'[83]of the digital era, is the primary business asset of these platforms.[84]They typically host user-generated content and data, track users' online behavior to generate profiles for targeted advertising, and either sell data to third parties or use it to develop products and services, including the training of GenAI models.[85]

Many digital platforms prohibit data scraping and other methods of automated data collection in their various contracts with users.[86]For instance, LinkedIn's user agreement prohibits users from making steps to "[d]evelop, support or use software, devices, scripts, robots or any other means or processes … to scrape the Services or otherwise copy profiles and other data from the Services."[87] Data scraping refers to "the process of extracting and combining content of interest from the Web in a systematic way…" and involves taking "raw data in the form of HTML code from sites and convert

80. *See infra* notes 86–98.

81. *See infra* notes 134–51.

82. *The World's Most Valuable Resource is no Longer Oil, But Data*, ECONOMIST (May 6, 2017), https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data (last visited Dec. 29, 2023).

83. Sandro Shubladze, *How to Make Use of the New Gold: Data*, FORBES (Mar. 27, 2023), https://www.forbes.com/sites/forbestechcouncil/2023/03/27/how-to-make-use-of-the-new-gold-data/?sh=1ad429042bbf (last visited Dec. 29, 2023).

84. ECONOMIST, *supra* note 82.

85. Data on digital platforms is essential for research in social science, in language, in medicine, and also for the purpose of studying and oversighting the social implications of platforms.

86. Niva Elkin-Koren, Maayan Perel & Ohad Somech, *Unlocking Platform Data for Research*, IN. L.J. (forthcoming 2024).

87. LINKEDIN, *User Agreement*, § 8.2(2) (Feb. 1, 2022), www.linkedin.com/legal/user-agreement.

it into a usable structured format."[88] Similarly, Section 3.2.3 of Facebook's ToS provides that: "You may not access or collect data from our Products using automated means (without our prior permission) or attempt to access data you do not have permission to access."[89]

Contractual bans on scraping have been invoked by platforms even when scraping was intended solely for non-competing research purposes,[90] targeting investigative journalists, academic researchers or civil society organizations seeking to enhance transparency.[91] For instance, in a recent lawsuit filed by X Corp. against the Center for Countering Digital Hate (CCDH), a non-profit organization that conducted research on the dissemination of hateful content on social media, X alleged that CCDH had intentionally and unlawfully scraped data from the X platform (formerly known as Twitter), thereby violating its ToS.[92] Rejecting X's claim, the court noted that "X Corp. has brought this case in order to punish CCDH for CCDH publications that criticized X Corp.—and perhaps in order to dissuade others who might wish to engage in such criticism."[93]

Contractual bans on scraping were enforced against entrepreneurs seeking to use data for developing new products and services. For instance, LinkedIn filed a lawsuit against hiQLabs, a data analytics company, which scraped LinkedIn publicly available profiles.[94] hiQLabs used this data to develop new analytic tools, such as apps predicting the risk of a talented employee being recruited by a competitor ('skill mapping'), or diagnosing skill gaps in organizations.[95] LinkedIn argued that extracting information using these tools

88. Domenico Trezza, *To Scrape or Not to Scrape, This is Dilemma: The Post-API Scenario and Implications on Digital Research*, FRONTIERS (Mar. 15, 2023), https://www.frontiersin.org/articles/10.3389/fsoc.2023.1145038/full.

89. Meta Platforms, Inc. v. Bright Data Ltd., No. 23-cv-00077-EMC, 2024 U.S. Dist. LEXIS 11913 (N.D. Cal. Jan. 23, 2024).

90. *See, e.g.*, Axel Bruns, *After the 'APIcalypse': Social Media Platforms and Their Fight Against Critical Scholarly Research*, 22 INFO. COMMC'N. & SOC'Y 1544 (2019).

91. For instance, in 2018, after the research group ProPublica refused to cease from using a web-scraping tool to monitor political advertisements on Facebook during the U.S. elections, Facebook implemented technical measures that effectively blocked ProPublica's scraping tool. *See* Jeremy B. Merrill & Ariana Tobin, *Facebook Moves to Block Ad Transparency Tools—Including Ours*, PROPUBLICA (Jan. 28, 2019), https://www.propublica.org/article/facebook-blocks-ad-transparency-tools.

92. X Corp. v. Ctr. for Countering Digit. Hate Ltd., No. 23-cv-03836-CRB, 2024 U.S. Dist. LEXIS 53042 (N.D. Cal. Mar. 25, 2024).

93. *Id.* at *22.

94. hiQ Labs, Inc. v. Linkedin Corp., 31 F.4th 1180 (9th Cir. 2022).

95. Javier Torre de Silva & López de Letona, *The Right to Scrape Data on the Internet: From the US Case hiQLabs, Inc. v. LinkedIn Corp. to the ChatGPT Scraping Cases: Differences Between US and EU Law*, 5 GLOB. PRIV. L. REV. 5 (2024), https://doi.org/10.54648/gplr2024001.

violated its User Agreement, which explicitly prohibited scraping or copying by automated methods.[96]

In another recent case, *Meta Platforms, Inc. v. Bright Data Ltd.*,[97] the court examined whether Bright Data, a company that searches, collects and sells publicly available data from various websites, including Facebook and Instagram, was breaching Facebook's and Instagram's ToS. Bright Data utilizes data collected from digital platforms to offer new types of access services, such as a *proxy network* that enables anonymous browsing, or a *Web Scraper Integrated Development Environment*, which assists programmers in developing their own search tools.[98] Meta alleged that scraping was contrary to some restrictions in the ToS and therefore amounts to a breach of contract.[99]

To determine whether such restrictions applied by platform's contract are preempted under the 1976 Copyright Act, courts apply a two-part test.[100] First, the court must examine whether the work that would be affected by the contractual claim comes within the subject matter of copyright, as specified by sections 102 and 103.[101] If so, the court moves to examine the second prong, to determine whether the rights asserted under the state law claim are equivalent to any of the exclusive right listed under section 106 of the Copyright Act.[102]

Applying the first prong of the copyright preemption test to raw data is challenging since raw data is not copyrightable. To be eligible for copyright protection, a work must be an original work of authorship, namely originated with the author, and fixed in a tangible medium.[103] Raw data ("wholly factual information not accompanied by any original written expression") is therefore excluded from copyright protection because it is "discovered" rather than "author[ed]."[104]The Copyright Act explicitly excludes in section 102(b) ideas,

96. *See* LINKEDIN, *supra* note 87.

97. *Meta Platforms, Inc. v. Bright Data Ltd.*, No. 23-cv-00077-EMC, 2024 U.S. Dist. LEXIS 11913 (N.D. Cal. Jan. 23, 2024).

98. *Id.*

99. Based on construing the ToS, the court concluded that the contractual terms that restrict scraping do not apply to Bright Data's logged-off scraping of publicly viewable data, and therefore denied Meta's motion for summary judgment.

100. MLGenius Holdings LLC v. Google LLC, No. 20-3113, 2022 U.S. App. LEXIS 6206, at *4 (2d Cir. Mar. 10, 2022), *cert. denied*, 143 S. Ct. 2658 (2023).

101. *See* Jackson v. Roberts, 972 F.3d 25, 42 (2d Cir. 2020).

102. *Genius*, 2022 U.S. App. LEXIS 6206, at *4.

103. 17 U.S.C. § 102(a). *See* Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 345 (1991) (noting that originality is "[t]he *sine qua non* of copyright").

104. *See Feist*, 499 U.S. at 345–46. Justice O'Connor stated, for a unanimous court, that facts are categorically excluded from copyright protection because they never originate with

procedures, processes, systems, methods of operation, concepts, principles or discoveries from copyright protection.[105]Likewise, knowledge, truths ascertained, conceptions and ideas are considered "free as the air to common use."[106] As famously elaborated by the Court in *Feist Publications v. Rural Telephone Service*, the idea-expression dichotomy in copyright law ensures that only original expressions of facts—but not facts in themselves—are protected by copyright.[107] While data is unprotected, it is clearly regulated by copyright law, which ensures it is not excluded from the public domain.

The second prong of the preemption test may easily apply to legal claims that seek to prohibit any reproduction of the data, including derivative modifications made to adapt it for AI training models.[108] However, since this claim stems from a contractual restriction in the ToS, courts will need to decide whether the contract amounts to an "extra element" necessary to avoid preemption.[109]

The Second Circuit's recent decision on preemption in *Universal Instruments Corp. v. Micro Sys. Eng'g, Inc.* took the approach that more than basic contractual privity is required to transform an otherwise equivalent claim into one that is qualitatively different from a copyright infringement claim.[110] The Second Circuit also followed this approach in the *ML Genius* case.[111] Genius argued that since its breach-of-contract claim requires it to plead "mutual assent and valid consideration", and "assert[] rights only against the contractual counterparty, not the public at large," its claim was qualitatively different from

---

the author but are "discovered" rather than "author[ed]." However, this principle does not cover "constructed" or "created" facts, which may reflect original contribution of their author and thus might be protected. *See, e.g.*, Jessica Litman, *The Public Domain*, 39 EMORY L.J. 965, 996 (1990) (arguing that facts "do not exist independently of the lenses through which they are viewed").

105. 17 U.S.C. § 102(b).

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

107. *Feist*, 499 U.S. at 348 (noting that "all facts—scientific, historical, biographical, and news of the day" may not be copyrighted and are part of the public domain available to every person).

108. *See, e.g.*, Nicola Lucchi, *ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems*, EURO. J. RISK REG. 1 (2023).

109. *See supra* notes 37–39 and accompanying text.

110. *See, e.g.*, Universal Instruments Corp. v. Micro Sys. Eng'g, Inc., 924 F.3d 32, 49 (2d Cir. 2019) (explaining that the parties' "contractual privity does nothing to change the fact that vindication of an exclusive right under the Copyright Act" asserted through a breach-of-contract claim "is preempted by the Copyright Act").

111. MLGenius Holdings LLC v. Google LLC, No. 20-3113, 2022 U.S. App. LEXIS 6206, at *7 (2d Cir. Mar. 10, 2022), *cert. denied*, 143 S. Ct. 2658 (2023).

a copyright claim, and therefore should not be preempted.[112] The Second Circuit disagreed, noting that Genius was effectively contemplating for "a per se rule that all breach of contract claims are exempt from preemption."[113] Such a rule, according to the court, "would be in tension with our precedent holding that the general scope inquiry is 'holistic.'"[114] Because the court found that Genius sought to protect a right equivalent to the exclusive rights secured by copyright, it concluded that the contractual claim was preempted.[115]

As discussed earlier, however, there exists a major circuit split regarding the applicability of the preemption doctrine to contractual restrictions.[116] This results in a high level of uncertainty regarding whether scraping should be governed by copyright law or left to the discretion of market players through private ordering.

In this contested context, the guiding principles outlined by Samuelson on the copyright/contract interface may provide additional clarity.

First, decisions concerning pervasive restrictions involving copyrights should not be decided by boilerplate contracts alone. Contractual bans on scraping in digital platforms' ToS effectively establish a "right against the world."[117] While ToS technically apply only to people who accept them, any potential users of the platform might be bound by them.[118] Additionally, there are often no alternative ways to obtain access to the data but through the platform. For instance, anyone seeking to access a post on Instagram must do so through its infrastructure, which is accessible only to signed-up users.[119]Consequently, by restricting reproduction of publicly available data hosted on their servers, platforms de facto secure themselves a proprietary right in that data. Such a restriction embedded in boilerplate ToS seeks to bind all users and therefore, if enforceable, resembles the in rem characteristics of a property right.[120]

---

112. *Id.* at *3.

113. *Id.* at *9.

114. *Id.*

115. *Id.*

116. *See Genius*, 2022 U.S. App. LEXIS 6206, *petition for cert. filed* 2022 WL 3227953 (U.S. Aug. 5, 2022) (No. 22-121).

117. Elkin-Koren, *The Limits of Freedom of Contract*, *supra* note 58, at 104.

118. *Id.* (explaining that "the introduction of new distribution technologies blurs the distinction between rights in personam and rights in rem. The availability of direct communication with users and the technical ability to prevent any unlicensed access by technological fencing facilitate a regime that is very similar in its nature to a property regime.").

119. *See, e.g.*, Linnea Laestadius & Alice Witt, *Instagram Revisited*, HANDBOOK OF SOC. MEDIA RSCH. METHODS, 581 (2022); Anna Lenhart, *A Vision for Regulatory Harmonization to Spur International Research*, LAWFARE (May 2, 2023).

120. *See supra* notes 55–59 and accompanying text.

Second, to identify contractual restrictions that may conflict with copyright, it is necessary to examine whether these provisions prohibit an action that copyright law would otherwise permit. When platforms' ToS ban the scraping of data that is publicly available, they aim to limit the copying of data–a use that is otherwise permitted by copyright law. Raw data, such as data extracted from users' profiles or the syntax of language extracted from users' posts, is not protected by copyright law. From a copyright perspective, raw data that is hosted by platforms and made publicly available should remain in the public domain. Copyright law aims to promote progress not only by grating exclusive rights to authors, but also by ensuring a robust public domain that fosters further creation. Therefore, raw data, which serves as building blocks for future creators, is deliberately excluded from copyright protection. This reflects a delicate balance "between the interests of authors . . . in the control and exploitation of their writings and discoveries on the one hand, and society's competing interest in the free flow of ideas, information, and commerce on the other hand . . . ."[121] When platforms attempt to assert exclusive rights over users' data hosted on their servers, they essentially disrupt the delicate balance between exclusivity and access, articulated by copyright policies.[122] Accordingly, contractual restrictions on the scraping of raw data in platforms ToS may conflict with copyright law where data is deliberately left in the public domain.

Finally, in deciding whether enforcing such contractual restrictions may undermine copyright policy, it is crucial to consider the implications on copyright policy. The enforceability of sweeping contractual prohibitions by digital platforms on publicly available data scraping might compromise the primary goals of copyright law. Digital platforms often serve as unique access point to data, which can be indispensable for basic research.[123] Under copyright law, scraping for research is often considered fair use due to the transformative nature of use, and because it does not threaten the platforms' core market.[124] Researchers essentially engage in "non-expressive" copying,[125] and do not intend to compete with the platform in the data market. Copyright policy may also seek to promote transformative use by entrepreneurs who seek

121. *See, e.g.*, Sony Corp. of Am. v. Universal City Studios, 464 U.S. 417, 429 (1984).

122. *See generally*, Robert A. Kreiss, *Accessibility and Commercialization in Copyright Theory*, 43 UCLA L. REV. 1, 2–4 (1995).

123. For example, it may be essential for detecting early indicators of imminent natural disasters, identifying markers for infectious disease outbreaks, or developing new research methodologies employing Artificial Intelligence.

124. Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 6 J. COPYRIGHT SOC'Y U.S.A. 291, 301–02 (2019).

125. Lemley & Casey, *supra* note 52, at n. 92.

to develop competing products and service, provided they only obtained access to non-protected data and ideas.[126] If GenAI developers were required to obtain a license for non-expressive copying of publicly available webpages, transaction costs would be prohibitive. Ensuring access to online publicly available data could also play important role in promoting freedom of speech. Scraping platforms' data by independent academic researchers and investigative journalists might be crucial for generating unbiased evidence to guide public oversight and inform public debates regarding the social ramifications of digital platforms.[127]

Overall, this analysis suggests that the determination of core copyright choices should not be left to contractual agreements but rather governed by copyright law. The legitimacy of scraping unprotected data and extracting unprotected ideas (patterns, weights) for training GenAI models should not depend on the drafting of contract restrictions by digital platforms. This concern regarding the power of platforms to unilaterally shape the norms of accessing data was highlighted by the Court of Appeals in *LinkedIn v. hiQ*: "Giving companies like LinkedIn free rein to decide, on any basis, who can collect and use data—data that the companies do not own, that they otherwise make publicly available to viewers, and that the companies themselves collect and use—risks the possible creation of information monopolies that would disserve the public interest."[128]

Leaving the issue of access to non-protected data to contract law may stifle the reuse of such data for developing new creative works, hinder the ability to learn from that data, and restrict free speech by limiting effective oversight of digital platforms.

Another disadvantage of contract law for regulating access to platforms' publicly available data, is the high level of uncertainty regarding restrictions on access. Various sites may include different restrictions, and courts' construction of these provisions may also vary. The recent decisions in *Bright Data* and *LinkedIn v. hiQ* demonstrate this issue. Following a long-running scraping litigation, the district court in the hiQLabs/LinkedIn dispute concluded that the data analytics hiQ has breached LinkedIn's contract by scraping and using scraped data.[129] In the lawsuit construing similar contractual

126. Sega Enterp. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1522–24 (9th Cir. 1992); Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596 (9th Cir. 2000).

127. Niva Elkin-Koren, *A2D in Digital Platforms: Bridging the Transatlantic Divide*, VERFBLOG (Feb. 23, 2024), https://verfassungsblog.de/a2d-for-researchers-in-digital-platforms/#:~:text=Bridging%20the%20Transatlantic%20Divide,these%20platforms%20encounter%20growing%20obstacles.

128. hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985 (9th Cir. 2019).

129. hiQ Labs, Inc. v. LinkedIn Corp., 639 F. Supp. 3d 994 (N.D. Cal. 2022).

restrictions in Meta's ToS, the court held that there was no breach of contract since there was no evidence that Bright Data had scraped non-publicly available data in violation of the contract.[130] Such uncertainty regarding the validity of contractual restrictions, their scope, and construction by courts may raise transaction costs involved in determining and acquiring access rights, thereby strengthening barriers to the use of data for further creation.

In another lawsuit against Bright Data, the district court finally addressed the preemption challenge in the context of data scraping.[131] Dismissing X's breach of contract claim against Bright Data, for scraping contrary to the ToS, the court held that: "invoking state contract and tort law, X Corp. would entrench its own private copyright system that rivals, even conflicts with, the actual copyright system enacted by Congress."[132] The court held that X's ToS that ban scraping conflict with copyright law since they prohibit the scraping of posts, which are not owned by X, but rather owned by its users. An enforceable contractual ban on scrapping would enable X Corp., which is a non-exclusive license, to exclude others from reproducing, adapting, distributing, and displaying works that belong to its users, thus conflicting with users' copyrights in the content, and also with the fair use privileges of third parties.[133] Moreover, the enforceability of such a contractual ban "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" thus shrinking the public domain.[134]

The court's analysis demonstrates how the focus on copyright policy, proposed by Samuelson, could help address the ban on scraping in platforms' ToS. Platforms' attempt to secure exclusive rights to publicly available data by contractually prohibiting scraping conflict with copyright policy and therefore should be preempted by copyright law.

C. RESTRICTING GENAI TRAINING BY EULA

Another context where contractual provisions related to copyrightable materials are invoked is where right holders rely on their EULA.[135] Recently,

130. Meta Platforms, Inc. v. Bright Data Ltd., No. 23-cv-00077-EMC, 2024 U.S. Dist. LEXIS 11913 (N.D. Cal. Jan. 23, 2024).

131. X Corp. v. Bright Data Ltd., No. C 23-03698 WHA, 2024 U.S. Dist. LEXIS 84708 (N.D. Cal. May 09, 2024).

132. *Id.* at *20.

133. *Id.* at *23–24.

134. *Id.* at *24.

135. EULAs were used early on by software providers to define the scope of protection for software, at a time when it was still unclear whether software is entitled to any intellectual

such breach of contract claims were raised in several pending class action lawsuits regarding the use of copyrighted materials in training GenAI systems.[136] In one lawsuit, the plaintiffs argued that Stable Diffusion, Midjourney, and DeviantArt infringed upon their copyrights by using images for training without proper authorization.[137] The plaintiffs allege that the models not only violated the license agreement but also infringed upon their copyright by copying images without authorization to generate derivative works.[138] Another lawsuit raised similar claims against OpenAI's flagship application ChatGPT.[139]

One of the most interesting noteworthy cases involving a breach of contract claim based on EULA is *Doe 1 v. GitHub, Inc.*[140] GitHub, a popular collaborative software development tool (acquired by Microsoft), offers public repositories for open source software.[141] On November 3, 2022, several unnamed software developers filed a class action lawsuit against GitHub, OpenAI, and Microsoft, concerning Copilot, a code editor extension which generates code suggestions in response to a users' prompt, and Codex, a generative AI model trained on publicly available code, including those in GitHub's repositories.[142] Plaintiffs contended that their Open Source Systems (OSS) stored on GitHub were used to train Copilot and Codex without authorization, and that their source code was subsequently reproduced in Copilot suggestions without proper attribution.[143] While the court dismissed

property protection. When copyright laws worldwide were expanded to cover computer programs, and in some jurisdictions, software was even granted patent protection, licenses were used for acquiring additional legal protection.

136. *See, e.g.*, Doe 1 v. GitHub, Inc., 672 F. Supp. 3d 837 (N.D. Cal. 2023); Authors Guild v. OpenAI Inc., No. 1:23-cv-08292 (S.D.N.Y. Sept. 19, 2023); Andersen v. Stability AI, Ltd., 700 F. Supp. 3d 853 (N.D. Cal. 2023); Getty Images v. Stability AI, No. 23-CV-00135 (D. Del. 2023); The N.Y. Times Co. v. Microsoft Corp., OpenAI, No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023).

137. *See* Andersen, 700 F. Supp. 3d at *1 (N.D. Cal. Oct. 30, 2023).

138. Specifically, they claimed that the model infringes copyrights because it was (1) trained on copyrighted code without authorization, and (2) generated snippets of that same copyrighted code. *Id.* at 21–24.

139. *See* Tremblay v. OpenAI, Inc., Nos. 23-cv-03223-AMO, 2024 U.S. Dist. LEXIS 110905, at *1, *11–12 (N.D. Cal. June 24, 2024).

140. Doe 1 v. GitHub, Inc., 672 F. Supp. 3d 837 (N.D. Cal. 2023).

141. According to public reports, it has more than 100 million registered users and hosts over 300 million code repositories. Back in 2018, Microsoft acquired GitHub for a reported $7.5 billion. GitHub's annual recurring revenue stands at $1 billion.

142. *See* Doe 1 v. GitHub, Inc., 672 F. Supp. 3d at 847 (N.D. Cal. 2023).

143. *Id.*

some claims in preliminary proceedings, it allowed the plaintiff's breach of OSS licenses claim to proceed.[144]

What makes this legal dispute particularly intriguing is that it concerns OSS licenses. Open source initiatives aim to enhance access to software.[145] Beginning with the General Public License (GPL) introduced by the Free Software Foundation (FSF) in the late 1980s, free software licenses seek to promote four freedoms pertaining to computer programs: the freedom to run the program, to study it, to modify it, and to share or redistribute both the original program and any modified versions.[146] The Open Source Initiative, established in 1998, pursued a more permissible approach, outlining essential criteria that a license must satisfy to qualify as open source, resulting in a plethora of open source licenses.[147]

In this lawsuit, the plaintiffs raised claims under both copyright law and California state law,[148] including a breach of contract claim. They alleged that defendants violated the OSS license agreement, which mandates attribution.[149]

144. Defendants, GitHub, OpenAI, and Microsoft, argued that (1) the plaintiffs were anonymous, (2) the plaintiffs did not provide specific instances where Copilot reproduced their licensed code, (3) GitHub's Terms of Service (TOS) grants broad rights to use and reproduce code. Moreover, the TOS preempts the breach of license claims. The court allowed the Digital Millennium Copyright Act (DMCA) violation, breach of OSS licenses, unjust enrichment, and unfair competition to proceed. On June 8, 2023, plaintiffs filed an amended complaint. This analysis is focused on plaintiffs' claims of violation of the DMCA and breach of open-source licenses governing their OSS.

145. *See* Niva Elkin-Koren, *What Contracts Cannot Do: The Limits of Private Ordering in Facilitating a Creative Commons*, 74 FORDHAM L. REV. 375, 391–92 (2005).

146. The GPL has a viral effect and applies automatically to any new copy of the software and any derivative program based on the original one. Subsequent creators and users are therefore bounded by the terms of access defined by the license and must strictly apply them to any subsequent work they create using the original free software.

147. *See* Elkin-Koren, *What Contracts Cannot Do*, *supra* note 145, at 411–13.

148. Several state law claims were dismissed in the preliminary proceedings, including tortious interference in a contractual relationship, fraud, false designation of origin, unjust enrichment, negligence, breach of the GitHub Privacy Policy and Terms of Service, violation of the California Consumer Privacy Act (CCPA) and unfair competition. *See* Doe v. GitHub, Inc., No. 22-cv-06823-JST, 2024 U.S. Dist. LEXIS 11068 (N.D. Cal. Jan. 3, 2024).

149. Plaintiffs argue that defendants failed to include in the output of the Copilot extensions: (1) attribution to the owner, (2) a copyright notice, and (3) the license terms, contrary to explicit provisions laid out in the OSS licensing terms. Doe 3 v. GitHub, Inc., No. 22-cv-7074-JST, ECF No. 1, 34 n. 4 (N.D. Cal. Nov 03, 2022). For instance, one Plaintiff in this class action (Doe 1) alleges that the relevant OSS license agreement (MIT License) requires attribution. Since Copilot allegedly uses the program in the training, and the output includes a verbatim copy of plaintiff's original code, such use constitutes both copyright infringement and a breach of contract. *See* Doe v. Github, Inc., No. 22-cv-06823-JST, 2024 U.S. Dist. LEXIS 11068, 861 (N.D. Cal. Jan. 3, 2024). This decision, however, does not include any findings on the breach of contract claim since the defendants did not move to dismiss

Allegedly, since Copilot utilized the code in the training, the output may include verbatim copies of plaintiff's original code, resulting in both copyright infringement and a breach of contract.

One approach to rebutting the contractual claims is to establish that GitHub was authorized to use software in its repository based on its ToS, and the license obtained from its users, including the plaintiffs, upon signing for the service.[150] GitHub's ToS grants it the right to "store, archive, parse, and display" copies of code uploaded by users. Furthermore, it is permitted to "parse it into a search index or otherwise analyze it" and publish the results for the use of others.[151] Users who opt to make their repositories public "grant each User of GitHub a nonexclusive, worldwide license to use, display, and perform" their content through the GitHub Service and to reproduce the content "solely on GitHub as permitted through GitHub's functionality."[152] Accordingly, if the use for training the model falls within the scope of the license acquired by GitHub, there is no copyright infringement.

The breach of contract claim, however, also involves a failure to fulfill an affirmative duty stipulated in the contract, to provide attribution in derivate works, arguably, in Copilot/Codex output.

Should courts enforce contractual obligations to provide attribution when software is used during training? Here again, the principles outlined by Samuelson may prove beneficial.

First, EULA, akin to ToS, resembles a boilerplate contract. Consequently, contractual obligations set forth by EULA may affect copyright policy. If rights holders can enforce duties against third parties, extending beyond the bundle of rights defined by copyright, they are effectively able to unilaterally establish a new quasi-copyright regime. When the legal obligations specified by the license become enforceable against all subsequent users, these users are compelled to ascertain which restrictions from the multitude of applicable licenses pertain to their respective use, merely to avoid infringement. This explains why the law may be hesitant to enforce contracts that run with the asset —such as an easement, which bind obligations directly to the asset rather than to the contracting parties—and why claims against third parties are often

Plaintiffs' claims for breach of contract for open-source license violations or breach of contract for selling licensed materials.

150. GitHub requires all of its users to agree to its terms of service. Doe 3 v. GitHub, Inc, No. 22-cv-06823-JST, ECF No. 108, 110 (N.D. Cal. Jan 22, 2024).

151. *Id.*

152. *See* Doe 3 v. GitHub, Inc., No. 22-cv-7074-JST, ECF No. 1, 27–28 (N.D. Cal. Nov 03, 2022). Pamela Samuelson, *Legal Challenges to Generative AI, Part I*, 66 COMM. OF THE ACM. 20 (July 2023).

dismissed.[153] These costs of compliance may inadvertently raise barriers to access and use creative works.[154]

The second issue is determining whether any contractual provisions may conflict with copyright. As noted by Samuelson and her coauthors, contractual obligations "that require action, such as giving attribution, that copyright law otherwise does not expressly require" may clash with copyright law.[155] An author's right to receive attribution in copyrighted materials is not enumerated under U.S. copyright law, and no such right applies to authors of computer programs.[156] The plaintiffs in the GitHub/Copilot case are asserting a new legal right, not granted by copyright, and therefore cannot be licensed.

On the other hand, open-source licenses typically mandate attribution. Although there are various types and versions of the OSS licenses, the vast majority clearly require providing attribution when code is used.[157] Indeed, compliance with OSS often comes down to crediting the rights holder and sharing the source code covered by the OSS license.

The breach of contract claim based on the OSS license could be interpreted as a conditional license, where the license requires attribution to the original author, and if the licensee fails to satisfy that condition the use of the code would be unauthorized, and therefore infringing. The Federal Circuit in *Jacobsen v. Katzer* held that an open source software, was subject to a conditional property license.[158] Consequently, the court held that incorporating an open source software into a commercial software, without meeting the conditions (proper attribution, reference to the license and the tracking of changes in the program) was an unlicensed modification and distribution of

153. *See generally*, Molly Shaffer Van Houweling, *Exhaustion and the Limits of Remote-Control Property*, 93 DENV. L. REV. 951 (2016); Molly Shaffer Van Houweling, *The New Servitudes*, 96 GEO. L.J. 885 (2008); Thomas W. Merrill & Henry E. Smith, *The Property/Contract Interface*, 101 COLUM. L. REV. 773 (2001) (arguing that property rights communicate a standard bundle of rights related to an asset explaining the objection to new forms of rights *in rem* in reducing information costs). Rights of exclusion, which are automatically imposed against third parties, may increase the information cost of potential users who seek to avoid copyright infringement. *See* Clarisa Long, *Information Costs in Patent and Copyright*, 90 VA. L. REV. 465, 482–83 (2004).

154. *See generally* Elkin-Koren, *What Contracts Cannot Do*, *supra* note 146.

155. *See* Samuelson, *The Copyright Principles Project*, *supra* note 8.

156. Indeed, Section 1202 of the Digital Millennium Copyright Act (DMCA) provides that Copyright Management Information (CMI) that is conveyed with the work should not be removed or altered by third parties. Recently the court dismissed Plaintiffs' claims under Section 1202(b)(1) and 1202(b)(3) of the DMCA with leave to amend. Doe 3 v. GitHub, Inc., No. 22-cv-06823-JST, ECF No. 108, 110 (N.D. Cal. Jan 22, 2024).

157. Doe 3 v. GitHub, Inc., No. 22-cv-06823-JST, ECF No. 50, 53 (N.D. Cal. Nov 05, 2023).

158. Jacobsen v. Katzer, 535 F.3d 1373 (Fed. Cir. 2008).

copyrighted materials, which constituted copyright infringement.[159] Yet, unlike more typical OSS legal disputes, where rights holders rely on their copyright in the code, it is unclear whether simply using code for training is a copyright infringement.[160]

Furthermore, the interpretation of compliance with the attribution requirement when using code to train GenAI models remains unclear. In *Jacobsen v. Katzer*, the Federal Circuit emphasized the importance of the OSS license agreement in fostering collaboration among numerous contributors and offering reassurance to downstream users against liability.[161] However, these considerations might lead to a different outcome when publicly available software is utilized for training GenAI systems.

While compliance with open-source attribution requirements is relatively straightforward, in the case of works used in training it is likely to be far more burdensome and ambiguous. Foundation models are trained on vast datasets capable of encompassing substantial portions of society's entire corpus of documented cultural knowledge.[162] This means hundreds of billions of words, or essentially the entire internet.[163] The extensive datasets upon which foundation models are built enable these models to excel in generalizing from seen to unseen examples.[164] This implies that models effectively deconstruct and reconstruct patterns: they "learn" concepts, principles, and generalities from their training datasets and then apply them to generate entirely new data that are logically, semantically, or phonetically similar (but not identical) to the

159. *Id.*

160. *See infra* notes 161–164 and accompanying text.

161. Jacobsen v. Katzer, 535 F.3d 1373 (Fed. Cir. 2008).

162. *See generally* Bommasani et al., *supra* note 67, at 131. For example, Google's T5 and Facebook's LLaMA datasets were trained on content from over 15 million websites. Similarly, OpenAI's GPT-3 features a model size of 175 billion parameters and was trained with 45 terabytes of data. Kevin Schaul, Szu Yu Chen & Nitasha Tiku, *Inside the Secret List of Websites that Make AI Like ChatGPT Sound Smart*, WASH. POST (Apr. 19, 2023). https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/. Similarly, OpenAI's GPT-3 features a model size of 175 billion parameters and was trained with 45 terabytes of data. *See* Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu & Lichao Sun, *A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT*, ARXIV (2023), https://arxiv.org/pdf/2302.09419.pdf.

163. Frederik Bussler, *Will the Latest AI Kill Coding?*, MEDIUM (July 21, 2020) https://towardsdatascience.com/will-gpt-3-kill-coding-630e4518c04d (last visited Feb. 17, 2024); Alex Hughes, *ChatGPT: Everything You Need to Know About OpenAI's GPT-4 Tool*, SCIENCE FOCUS (Sept. 25, 2023), https://www.sciencefocus.com/future-technology/gpt-3 (last visited Dec. 31, 2023).

164. Generalization refers to the ability to adapt to unseen examples drawn from seen data distribution.

data in their training datasets.[165] The contribution of each particular piece of work to this process of generalization is often untraceable. Consequently, it might not be feasible to impose an obligation to provide attribution that would be enforceable against all downstream users of such models.

Finally, to determine whether contractual restrictions should be overridden by copyright, it is necessary to ascertain whether enforcing them would conflict with copyright policy. Granting quasi-copyright protection to the attribution requirement, which extends beyond the exclusive rights granted under copyright law, may clash with copyright objectives. Such an approach is likely to create a contractual thicket with conflicting provisions, resulting in unnecessary barriers to potential innovative uses of code in the GenAI ecosystem. This would undermine the balance struck by copyright law between providing incentives to authors and facilitating access to works—a balance essential for promoting progress. Enforcing duties to provide attribution in OSS against the use in GenAI training, may therefore impede the advancement of functional learning from existing code, conflicting not only with the objectives of copyright law but also with the purpose of the open-source movement.

IV. CONCLUSIONS

Despite its rapid and widespread penetration into various aspects of our lives, GenAI is still in its infancy development stage, making it difficult to predict the full scope of social consequences it will carry.[166] Consequently, there is relatively little GenAI related legislation, and regulatory expertise is still developing. This may reinforce the emergence of private ordering solutions to address some of the contested issues as described in this paper. However, it does not necessarily mean that these private ordering arrangements will be more efficient or considerate of the implications for social welfare.

Contract law may facilitate experimentation with different allocations and contribute to the development of new types of transactions between stakeholders, thus supporting innovative business models. However, boilerplate contracts involving creative works should be carefully examined through the lens of copyright policy. Prohibitions on scraping in the ToS of digital platforms or new obligations created by contracts, which are not otherwise mandated by copyright law, should be examined in light of copyright principles. Unlike contract law, copyright law embodies important checks and

165. *See, e.g.*, Andrew Brock, Jeff Donahue & Karen Simonyan, *Large Scale GAN Training for High Fidelity Natural Image Synthesis*, ARXIV (2018), https://arxiv.org/abs/1809.11096.

166. *See generally*, DAVID COLLINGRIDGE, THE SOCIAL CONTROL OF TECHNOLOGY (1980).

balances enforced by courts. Courts would need to determine whether scraping bans involve the legitimate reproduction of unprotected facts or are otherwise permissible under the fair use doctrine.[167] Similarly, courts could evaluate how new obligations created by boilerplate contracts impact the objectives of copyright law. Even if such provisions are not entirely preempted, courts could interpret those that conflict with copyright policy more narrowly.[168]

Samuelson was aware of the complexity raised by the copyright/contract interface and therefore opposed a categorical ban that would render unenforceable any contract terms contradicting copyright.[169] Instead, she advocated for incremental development of the law, whereby standards for enforceability of such restrictive terms would evolve through court rulings.[170] This incremental approach is useful for facilitating trial and error not only in developing legal norms but also in shaping new emerging technologies and infrastructures.

167. The Copyright Act § 107 ("[T]he fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright.").

168. *See* Elkin-Koren, et al., *supra* note 86.

169. Samuelson, *Possible Futures of Fair Use Symposium*, *supra* note 7, at 859–60.

170. *Id.*