COPYRIGHT IN AI TRAINING DATA: A HUMAN-CENTERED APPROACH

DAVID W. OPDERBECK[*]

Table of Contents

[*] Professor of Law and Co-Director, Gibbons Institute of Law, Science & Technology and Institute for Privacy Protection, Seton Hall University School of Law. Thanks to Ben Sobel, Matthew Sag, Pam Samuelson, Michael Carroll, Michael Madison, Chris Newman, Courtney Cox, Lateef Mtima, Jon Romberg, the participants at the Lastowka Intellectual Property Colloquium, and the participants at the Colloquium on Philosophical Methods in Intellectual Property for valuable comments on earlier drafts of this paper.

I. Introduction

AI systems require training. AI training requires large volumes of examples. The examples used to train AI systems, siphoned from the public Internet, often are subject to copyrights. This massive unlicensed use of copyrighted material implicates the reproduction right because these systems must make copies of files to analyze them. It also implicates the right to control derivative works to the extent the trained system is "based upon" the training data.[1]

In May 2023, the U.S. Copyright Office held listening sessions on AI and the visual arts that focused on the use of copyrighted works in training data.[2] The Federal Trade Commission ("FTC") issued an investigative demand to OpenAI, the creator of the text generator ChatGPT, that included requests for information about the sources of its training data.[3] The European Union ("EU") is considering rules that would require disclosure of copyrighted material used in training data.[4]

Groups of authors and other content creators have filed lawsuits against OpenAI for ingesting their content without permission to train large language models.[5] These includes a class action brought by The Authors

[1] *See* 17 U.S.C. §§ 101, 106.

[2] U.S. Copyright Off., Libr. of Cong., Transcript of Proceedings: Copyright on Artificial Intelligence and Visual Arts Listening Session 1 (May 2, 2023), https://www.copyright.gov/ai/transcripts/230502-Copyright-on-AI-and-Visual-Arts-Listening-Session-revised.pdf [hereinafter Transcript of Proceedings].

[3] Fed. Trade Comm'n, FTC File No. 232-3044, FTC Civil Investigative Demand ("CID") Schedule at 5 (n.d.), https://www.washingtonpost.com/documents/67a7081c-c770-4f05-a39e-9d02117e50e8.pdf?itid=lk_inline_manual_4 (section II.A.15, under "Interrogatories"); *id.* at 18-19 (sections II.B.7, II.B.13, under "Reasons for Documents").

[4] Supantha Mukherjee et al., *EU Proposes New Copyright Rules for Generative AI*, REUTERS (Apr. 28, 2023, 1:51 AM), https://www.reuters.com/technology/eu-lawmakers-committee-reaches-deal-artificial-intelligence-act-2023-04-27/.

[5] *See* Jonah Valdez, *Sara Silverman and Other Bestselling Authors Sue Meta and OpenAI for Copyright Infringement*, L.A. TIMES (July 10, 2023, 5:03 PM), https://www.

Guild, with prominent authors such as John Grisham and George R.R. Martin joined as individual plaintiffs.[6] Notably, in October 2023, Google promised to defend and indemnify users of its AI platforms against copyright claims.[7] Other lawsuits, as well as regulatory and legislative inquiries involving the use of copyrighted material for AI training, will certainly follow. Indeed, this issue is the next great frontier in copyright law, which will shape both the law and this revolutionary technology as much as the dawn of the computer and Internet eras did over forty years ago.

The training and deployment of this first wave of AI systems mirrors earlier Silicon Valley culture: move fast, break things, ignore intellectual property rights and ethical conundrums, and sort out the problems later. This pattern is etched deeply into intellectual property law and scholarship. The 1980s saw cases involving arcade video games and personal computers; the 1990s and early 2000s saw policy choices about the Internet and cases concerning peer-to-peer file sharing and the digitization of newspaper archives; the mid-2000s saw litigation over the Google Books project (also spearheaded by The Authors Guild), cable television, and cloud-based DVRs; and the early 2020s saw disputes about operating system Application Programming Interfaces ("APIs").[8]

Some scholars argue that the arc of intellectual property law over the past forty years bends towards fair use.[9] They envision a broad fair use

latimes.com/entertainment-arts/books/story/2023-07-10/sarah-silverman-authors-sue-meta-openai-chatgpt-copyright-infringement; Blake Brittain, *Lawsuit Says OpenAI Violated US Authors' Copyrights to Train AI Chatbot*, REUTERS (June 29, 2023, 1:55 PM), https://www.reuters.com/legal/lawsuit-says-openai-violated-us-authors-copyrights-train-ai-chatbot-2023-06-29/; Class Action Complaint, Doe v. Github, Inc., No. 3:22-cv-06823 (N.D. Cal. Nov. 3, 2022); Andersen v. Stability AI LTD., No. 3:23-cv-00201 (N.D. Cal. Jan. 1, 2023); Complaint at 1, Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135 (D. Del. Feb. 3, 2023); Class Action Complaint, Silverman v. OpenAI, Inc., No. 3:23-cv-03416 (N.D. Cal. July 7, 2023); Kadrey v. Meta Platforms, Inc., No. 3:23-cv-03417, 2023 WL 8039640 (N.D. Cal. 2023).

[6] Press Release, Authors Guild, The Authors Guild, John Grisham, Jodi Picoult, David Baldacci, George R.R. Martin, and 13 Other Authors File Class-Action Suit Against OpenAI (Sept. 20, 2023), https://authorsguild.org/news/ag-and-authors-file-class-action-suit-against-openai/.

[7] Neal Suggs & Phil Venables, *Shared Fate: Protecting Customers with Generative AI Indemnification*, GOOGLE CLOUD (Oct. 12, 2023), https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification.

[8] *See infra* Parts III, IV.

[9] *See infra* Part IV.

domain for "non-expressive uses" to accommodate disruptive technologies.[10]

There are some important problems with this vision. First, the arc of fair use bends in various, sometimes inscrutable ways. It is unclear whether any sort of non-expressive use principle can be gleaned from computer-age case law. It is even less clear whether such a principle would be doctrinally and practically coherent.

Second, the AI revolution is different, both in scale and in ethical concerns. In the 1990s, people were amazed that a computer hard drive could hold hundreds of songs and that an entire album could reside on a portable MP3 player.[11] In the early 2000s, people were astonished that researchers could find digital copies of old newspaper articles on the NEXIS database rather than by rummaging through microfiche.[12] By the mid-2000s, people were awed that Google could scan over twenty million library books.[13] Today, large language model ("LLM") AIs such as ChatGPT consume *billions* of files for training purposes, including publicly accessible songs, newspaper articles, books—and much, much more.[14] Technologists predict that advances in storage, communications, and computing power will allow the next generations of AIs to make equally impressive leaps in scale.[15]

So far, much of the scholarship on intellectual property rights in AI training data assumes that AI presents the same doctrinal and ethical concerns as previous generations of digital era technologies.[16] Many scholars' arguments appear rooted in a prior generation's computer and Internet exceptionalism.[17] Intellectual property rights, they suggest, are barriers on the road to greater knowledge and cultural diffusion.[18]

[10] *See infra* Part IV.

[11] *See* Daniel Ionescu, *Evolution of the MP3 Player*, PCWORLD (Oct. 29, 2009, 5:45 PM), https://www.pcworld.com/article/520590/evolution_of_the_mp3_player.html.

[12] *See infra* Section III.B.

[13] *See* Stephen Heyman, *Google Books: A Complex and Controversial Experiment*, N.Y. TIMES (Oct. 28, 2015), https://www.nytimes.com/2015/10/29/arts/international/google-books-a-complex-and-controversial-experiment.html.

[14] *See infra* Part II.

[15] *See* ERIK BRYNJOLFSSON & ANDREW MCAFEE, THE SECOND MACHINE AGE: WORK, PROGRESS, AND PROSPERITY IN A TIME OF BRILLIANT TECHNOLOGIES 13-38 (2014).

[16] *See infra* Part IV.

[17] *See infra* Part IV.

[18] *See infra* Part IV.

AI ethics scholars and policymakers are not so sanguine.[19] Beyond mere scale, AI's logarithmic growth raises ethical questions and deeper, lurking issues. MP3 players and scanned library books do not make decisions that affect people's lives and freedoms. AIs do. Nor is there any debate about whether an MP3 player or a page scan possesses legal rights of its own. AIs might. Perhaps we learned something from the hubris of the Internet age, which produced both enormous, glorious cultural goods and grave, corrosive evils.

Much of the work that has been done on AI ethics, policy, and law focuses on what AI knows and the decisions it makes about human beings. The Biden Administration's "Blueprint for an AI Bill of Rights," for example, emphasizes safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation, as well as human alternatives, consideration, and feedback.[20] The current draft of an EU AI Regulation reflects similar concerns.[21] The Future of Life Institute's Asilomar AI principles include broad statements about "shared benefit" along with the usual concerns around safety, transparency, and accountability.[22] Notably, none of these policy documents suggest principles for intellectual property.[23]

This presents an important opportunity for copyright to make a difference. Four aspects of copyright doctrine intersect with AI ethics in interesting ways. The first intersection is the meaning of reproduction. Copyright law considers any fixation in a tangible medium of expression sufficient both for purposes of obtaining statutory copyright and for purposes of defining a "copy" under the right of reproduction.[24] Undoubtedly, a reproduction is made of AI training data until the machine incorporates that data into its algorithmic functions.[25] This process could be considered transitory in a way that does not infringe the reproduction

[19] *See, e.g.*, *Blueprint for an AI Bill of Rights*, THE WHITE HOUSE, https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (last visited Feb. 15, 2024).

[20] *Id.*

[21] *See* Mukherjee et al., *supra* note 4. The Responsible AI Global Policy Framework published by ITechLaw includes a section on protecting intellectual property generated by AI but says nothing about intellectual property consumed by an AI. Susan Barty et al., *AI and Intellectual Property*, *in* ITECHLAW, RESPONSIBLE AI: A GLOBAL POLICY FRAMEWORK 257 (Charles Morgan ed., 1st ed. 2019), https://www.itechlaw.org/ResponsibleAI.

[22] *Asilomar AI Principles*, FUTURE OF LIFE INST. (Aug. 11, 2017), https://futureoflife.org/open-letter/ai-principles/.

[23] Barty et al., *supra* note 21.

[24] 17 U.S.C. § 102.

[25] *See infra* Section III.A.

right.[26] But the underlying data does, in a sense, survive in the algorithmic functions. In many ways, this is similar to how the human brain processes and recalls information, as the moniker "neural network" suggests.

The second intersection is consent. Licensed use of copyrighted works, of course, is not infringement.[27] This means a copyright owner can consent to a use of their copyright, either expressly or impliedly. Current scholarship on copyright and AI training data assumes that the basis for earlier examples of large-scale web crawling and scraping—notably Internet searches—is fair use and that fair use therefore must also be the primary basis for using AI training data. This is not so: the more prosaic rationale is consent through express or implied licenses.[28]

Consent is also a central pillar of AI ethics, particularly as those ethics intersect with privacy law. This includes consent to be subject to automated decision-making and consent to the processing of personally identifiable information ("PII") by an AI.[29] This pillar of AI ethics demonstrates the close connection between AI ethics and privacy law, which is grounded in basic human rights principles. Many kinds of texts and images that AIs ingest for training contain PII. The convergence of consent in both copyright and AI ethics suggests that more robust consent mechanisms for web crawling and scraping, supported by application design principles, would go a long way toward addressing future concerns about AI training and copyright along with related concerns about privacy.[30]

The third intersection relates to what seems to be the principal fair use defense raised by organizations such as OpenAI: non-expressive use. Many scholars and advocates assume non-expressive uses are inherently transformative.[31] They suggest that an open source ethic applicable to computer code, scientific findings, or discreet factual data held in databases maps directly on to AI training data. It does not.

Efficient computer code depends on good code and the progress of science depends on complete and accurate facts. Open source computer projects entail communities that vet and correct the code.[32] Scientific

[26] *See infra* Section III.B.

[27] *See* 17 U.S.C. § 201(d).

[28] *Cf.* Katherine Lee et al., *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*, J. COPYRIGHT SOC'Y (forthcoming 2024) (manuscript at 108-11), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4523551.

[29] *See Blueprint for an AI Bill of Rights*, *supra* note 19.

[30] *See infra* Section IV.C.

[31] *See infra* Part IV.

[32] *See* David W. Opderbeck, *The Penguin's Genome, or Coase and Open Source Biotechnology*, 18 HARVARD J.L. & TECH. 167, 192-93 (2004).

communities, to whatever degree they are open or closed source, depend on scientific methods, community norms, and peer review to weed out inaccurate data and conclusions.[33]

Presently, AI training is a wild west. AI models are trained from petabytes of data scraped from the Internet, and no one knows whether that data is good, bad, or indifferent. Copyright cannot serve as a primary mechanism for training data integrity, but it can serve as a useful speed bump. Even more, a market for clearing copyrighted content as AI training data would serve the purpose of copyright by benefitting content creators and enhance the integrity of training data through market forces.

The final intersection between copyright law and AI ethics is the value of education. "[E]ducational purposes" are mentioned in the Copyright Act of 1976 as an example of uses that could be fair under the "purpose and character of the use" factor.[34] Of course, educational uses are not per se fair uses, particularly when there is an established market for the kind of educational content at issue, but there are good reasons why educational uses are specifically mentioned in the statute. So, how does the value of education relate to machine learning? This question surfaces debates about the function of AI machines in human society, including whether an AI *itself* can have legal rights. A few scholars have considered whether an AI could possess rights as an author or inventor of things produced by the AI. But no one is asking whether an AI has a right to education that might factor into a fair use analysis of copyrighted training data, or at least whether AI machines' ability to educate—or miseducate—humans provides fodder for a fair use analysis.

Part II of this Article briefly reviews what AI is and how it learns. Part III discusses why AI training implicates the reproduction right. This involves a careful distinction, which the existing literature has rarely made, between the materials initially used to train an AI and the mathematical tokens stored within an AI. Part IV examines whether and to what extent a doctrine of non-expressive use should apply to the use of copyrighted materials for AI training. Part V turns to the novel question of education in the fair use analysis of AI training and explores themes in the emerging field of machine ethics regarding the rights of AI systems. Part VI concludes.

[33] *See* David W. Opderbeck, *A Virtue-Centered Approach to the Biotechnology Commons (or, The Virtuous Penguin)*, 59 ME. L. REV. 315, 329-30 (2007).

[34] 17 U.S.C. § 107.

II. AI Training, Reproduction, and Consent

A. What Is AI and How Does It Learn?

Early scholarship on artificial agents tended to blur distinctions among existing and potential types of agents.[35] Some of the important early scholarship focuses on what today we call "strong" AI or "artificial general intelligence" ("AGI").[36] AGI is an artificial agent with capacities for reason and awareness that equal or exceed human capacities.[37] As far as we know, AGI does not yet exist.[38] Some researchers and philosophers think AGI is both possible and likely while others believe there is something about the relationship between mind and body that makes AGI based solely in machines impossible.[39]

The kinds of AI we presently encounter, and those that will transform our lives in the future, are forms of "weak" or "narrow" AI—or, more accurately, forms of machine learning ("ML"), including LLMs such as ChatGPT.[40] ML systems use algorithms to process large amounts of data. Many ML systems are based on "neural networks," which roughly model how the human brain functions.[41] An "input" layer takes information from the outside world; this information is processed within "hidden" layers, which in "deep" neural networks may include millions of nodes (artificial neurons); and the final result of this process is communicated through an "output" layer.[42]

Advances in data storage, computing power, and network design enable vast nodal structures, each containing small data portions, with numerous potential pathways for an input to be analyzed before an output is produced. Like the human brain, the algorithms include parameters that allow these systems to "learn" as more and more data is processed, adjusting the

[35] *See* Lawrence B. Solum, Essay, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231, 1240-55 (1992).

[36] *See id.*; Reece Rogers, *What's AGI, and Why Are AI Experts Skeptical?*, WIRED (Apr. 20, 2023, 7:00 AM), https://www.wired.com/story/what-is-artificial-general-intelligence-agi-explained/.

[37] *See* Rogers, *supra* note 36.

[38] *See id.*

[39] *See* David W. Opderbeck, *Artificial Intelligence, Rights, and the Virtues*, 60 WASHBURN L.J. 445, 445-46 (2021) [hereinafter Opderbeck, *Artificial Intelligence*].

[40] *See What Is Machine Learning?*, IBM, https://www.ibm.com/topics/machine-learning (last visited Feb. 15, 2024).

[41] *See What Is a Neural Network?*, AWS, https://aws.amazon.com/what-is/neural-network/ (last visited Feb. 15, 2024).

[42] *Id.*

algorithmic weights in various nodes.[43] The small data portions retained by an ML system are not actual portions of the input training layer itself. Rather, the input layer is decomposed and translated into algorithmic representations that can be thought of as mathematical "tokens."[44]

Image recognition is a well-established and easily-understood application of this technology. Consider this digital photograph of a beach:[45]

Most people could immediately identify this photo as a "beach" scene. A little experience with actual beaches, or even with photos and videos of beaches, creates pattern recognition pathways in the brain.[46] If the image includes certain proportions, shapes, colors, and intensities—a bit of sky blue, a bit of ocean blue, a bit of sandy brown, a bit of green, all following

[43] *See* Transcript of Proceedings, *supra* note 2, at 35 (statement of Curt Levey) ("The trained model, consisting of millions or billions of weights, analogous to the synaptic connections in the human brain, retains no copies of the training examples.").

[44] *See* Lee et al., *supra* note 28 (manuscript at 6-48); Pamela Samuelson, *Generative AI Meets Copyright*, 381 SCIENCE 158, 159 (2023); Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295, 314-25 (2023); Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 57-59 (2017).

[45] Photograph of beach licensed from Adobe Stock.

[46] *See How the Brain Recognizes What the Eye Sees*, SALK INST. FOR BIOLOGICAL STUD. (June 8, 2017), https://www.salk.edu/news-release/brain-recognizes-eye-sees/.

something like the rule of thirds—there is a high probability the scene is a "beach" and not, say, a law school classroom.[47]

An ML image recognition system can mimic this process using the pixel data in a digital photo.[48] A typical medium-quality cell phone image contains millions of individual pixels, each with values for screen location, color, and intensity.[49] Groups of pixels with related location, color, and intensity can be assigned new algorithmic values, groups of groups can be assigned further values, and so-on, until the millions of individual pixels in the image are reduced to a small set of values that can be compared to algorithmic values from other photos.[50] With enough training data, the system can probabilistically distinguish "beach" photos from "classroom" photos quickly and accurately.

B. Crawling and Scraping

The process of AI training using publicly accessible data involves web crawling and web scraping. A web crawler is a program, often called a bot, that analyzes the code on a target website to create an index.[51] To do this, the crawler must at least make a temporary copy of the target website's code. Bots create indexes that can be used for various purposes, including searches.[52] A web scraper not only indexes information but also retrieves and stores content, such as text and images, from the target website.[53]

[47] *See id.*

[48] *See* Kinza Yasar, *Image Recognition*, TECHTARGET, https://www.techtarget.com/searchenterpriseai/definition/image-recognition?Offer=abt_pubpro_AI-Insider (last updated Mar. 2023).

[49] Digital camera sensors are measured in megapixels. *See* Pye Jirsa, *What Are Megapixels and Why Do They Matter?*, SLR LOUNGE, https://www.slrlounge.com/what-are-megapixels-and-do-they-matter-minute-photography/ (last visited Feb. 15, 2024). A megapixel is one million pixels. *Id.* Thus, a fifty-megapixel cell phone camera, such as that in the current Google Pixel 7 phones, produces images of fifty million pixels. *Id.*; *Pixel Phone Hardware Tech Specs: Pixel 7 Phones (2022)*, GOOGLE: PIXEL PHONE HELP, https://support.google.com/pixelphone/answer/7158570?visit_id=638436126758575445-271822901&p=specs&rd=1 (last visited Feb. 15, 2024).

[50] *See* Yasar, *supra* note 48.

[51] *See* Cem Dilmegani, *Web Crawler: What It Is, How It Works & Applications in 2024*, AIMULTIPLE: RSCH. (Jan. 10, 2024), https://research.aimultiple.com/web-crawler/.

[52] *Id.*

[53] *See* Cem Dilmegani, *In Depth Guide to Web Scraping for Machine Learning in 2024*, AIMULTIPLE: RSCH. (Jan. 2, 2024), https://research.aimultiple.com/machine-learning-web-scraping/.

Web crawling and scraping tools are readily available and easy to use.[54] There are also repositories of crawled and scraped web data available to anyone. One such repository, Common Crawl, boasts that it "contains petabytes of data collected since 2008," including "raw web page data, extracted metadata and text extractions," and other datasets available on the Amazon Web Services ("AWS") Data Exchange.[55] Other crawl datasets, such as LAION, include image and "aesthetic" data derived from Common Crawl data.[56] One of LAION's "Openclip" datasets contains 5.8 billion text-and-image pairs.[57] ChatGPT, the LLM text model that has generated so much excitement and concern, was trained in part on Common Crawl data.[58] DALL-E, the image-generation tool that generated comparable buzz, was partly trained on LAION data.[59]

C. Existing and Potential Markets for AI Training Data

In addition to open source public databases such as Common Crawl and LAION, there are burgeoning sources of academic and commercial training data derived from various sources, including the open Internet, the Dark Web, experiments, crowdsourcing, proprietary information, and partially synthetic and synthetic data.[60] The "Argoverse" data set, for example, includes data collected by researchers at Carnegie Mellon University and

[54] *See, e.g.*, stlane, *Lucene Website Crawler and Indexer*, CODE PROJECT (Jan. 31, 2009), https://www.codeproject.com/Articles/32920/Lucene-Website-Crawler-and-Indexer; ZYTE, https://zyte.com (last visited Feb. 15, 2024); *Documentation*, OPENSOLR, https://opensolr.com/faq/view/web-crawler (last visited Feb. 15, 2024).

[55] *See So You're Ready to Get Started*, COMMON CRAWL, https://web.archive.org/web/20230522201837/https://commoncrawl.org/the-data/get-started/ (last visited Mar. 16, 2024); *AWS Data Exchange*, AMAZON, https://aws.amazon.com/data-exchange/?adx-cards2.sort-by=item.additionalFields.eventDate&adx-cards2.sort-order=desc (last visited Apr. 14, 2024). Curiously, since an earlier draft of this article, Common Crawl removed this reference from its public website, perhaps in light of pending copyright litigation.

[56] *See* Romain Beaumont, *LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets*, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b/; Christoph Schuhmann, *LAION-Aesthetics*, LAION (Aug. 16, 2022), https://laion.ai/blog/laion-aesthetics/.

[57] Romain Beaumont, *Large Scale Openclip: L/14, H/14 and G/14 Trained on Laion-2B*, LAION (Sept. 15, 2022), https://laion.ai/blog/large-openclip/; *see also About ImageNet*, IMAGENET, https://image-net.org/about.php (last visited Feb. 15, 2024).

[58] Dennis Layton, *ChatGPT – Show Me the Data Sources*, MEDIUM (Jan. 30, 2023), https://medium.com/@dlaytonj2/chatgpt-show-me-the-data-sources-11e9433d57e8.

[59] *Id.*

[60] *See* Benjamin Sobel, *A Taxonomy of Training Data: Disentangling the Mismatched Rights, Remedies, and Rationales for Restricting Machine Learning*, *in* ARTIFICIAL INTELLIGENCE AND INTELLECTUAL PROPERTY 221, 229-36 (Jyh-An Lee et al. eds., 2021).

Georgia Institute of Technology using a fleet of autonomous vehicles.[61] As another example, the "Unsupervised Llamas" data set, provided by the German appliance maker Bosch, includes lidar-mapped lane markers, which are also used for training autonomous driving systems.[62]

There are also commercial providers of AI training data. Some of these providers obtain training data from their own Internet scrapes and from databases such as Common Crawl.[63] Many of these providers add value to other databases by structuring datasets—that is, by adding tags and other metadata so that the data is more useful and comprehensible from the start.[64] Yet others use a crowdsourcing model to obtain base data from individual contributors.[65] A crowdsourcing model can be useful to train language models on languages other than English.[66]

Other providers specialize in synthetic training data.[67] This can include both partially synthetic data, which creates datasets based on deidentified or otherwise modified real data (whether open or proprietary), and fully synthetic data, which are not derived from any real dataset.[68] As Michal Gal and Orla Lynskey note, synthetic data markets can improve the quality of training data, protect privacy, and enhance competition by reducing barriers

[61] *See* Ming-Fang Chang et al., *Argoverse: 3D Tracking and Forecasting with Rich Maps*, CORNELL UNIV. (Nov. 6, 2019), https://arxiv.org/pdf/1911.02620.pdf.

[62] *See Unsupervised Llamas: The Unsupervised Labeled Lane Markers Dataset*, BOSCH, https://unsupervised-llamas.com/llamas/ (last visited Feb. 15, 2024).

[63] *See, e.g.*, WEBZ.IO, https://webz.io/ (last visited Feb. 15, 2024) (claiming to provide "the world's largest structured web data feeds from across the open, deep, and dark web"); SCALE AI, https://scale.com/ (last visited Feb. 15, 2024); *How Does It Work?*, WEB.IO, https://docs.webz.io/reference/how-it-works-video (last visited Feb. 15, 2024).

[64] *See* Daniel Lee*, How We Scale Machine Learning,* SCALE AI (May 18, 2020), https://scale.com/blog/how-we-scale-machine-learning; *How Does It Work?*, *supra* note 63; *About Us*, COGITO, https://www.cogitotech.com/about-us/ (last visited Feb. 15, 2024); *About Us*, ANOLYTICS, https://www.anolytics.ai/ (last visited Feb. 15, 2024); WISEPL, https://www.wisepl.com/ (last visited Feb. 15, 2024); SUPERANNOTATE, https://www.superannotate.com/ (last visited Jan. 25, 2024).

[65] *See, e.g.*, *AI Requires a Human Touch: How Appen Recruits Crowds to Improve Technology*, APPEN (Aug. 31, 2017), https://appen.com/blog/ai-requires-human-touch-appen-crowd-recruiting/; *Introducing Defined.ai*, DEFINED AI (Oct. 25, 2021), https://www.defined.ai/blog/introducing-defined-ai/.

[66] *See AI Requires a Human Touch: How Appen Recruits Crowds to Improve Technology*, *supra* note 65.

[67] *See, e.g.*, SUPERANNOTATE, *supra* note 64.

[68] *See* Michal S. Gal & Orla Lynskey, *Synthetic Data: Legal Implications of the Data-Generation Revolution*, 109 IOWA L. REV. (forthcoming 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4414385.

to entry into markets that create AI products.[69] Gal and Lynskey note that sixty percent of AI training data will be synthetic by 2024.[70] As this survey suggests, markets for unstructured and structured AI training data are developing rapidly alongside the growth of AI use case and applications.

III. Initial Copyright Issues: Copying, Consent and Transitory Reproduction

A. Copying

The first question raised by AI training data is whether it involves copying at all. As discussed, an AI does not retain complete or partial copies of its training data.[71] Rather, it uses the training data to generate algorithmic tokens, which are employed within its multitude of artificial neurons to make probabilistic decisions.

Some commentators seem to suggest that AI training therefore might not implicate the reproduction right.[72] Pam Samuelson, for example, notes that because of the idea/expression dichotomy, "[p]hotographs of cats . . . do not give the photographer exclusive rights to characteristic features of cats, such as their noses or facial expressions."[73] Samuelson is, of course, correct about the features of cats. If the only portion of the cat photograph reproduced during training were the eyes, nose, and mouth, perhaps this would not constitute a reproduction. But this is not ordinarily how AI training data works.

[69] *Id.* (manuscript at 20, 28-29).

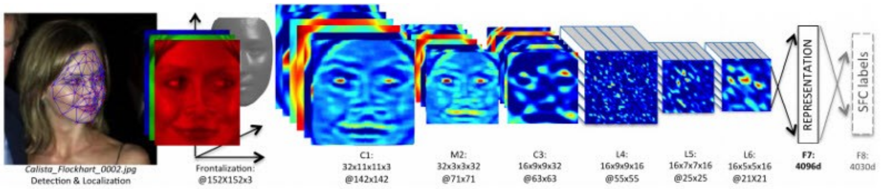
[70] *Id.* (manuscript at 3).

[71] *See supra* Section II.A.

[72] *See* Samuelson, *supra* note 44, at 159, 161; Sag, *supra* note 44, at 311-13.

[73] Samuelson, *supra* note 44, at 159.

A facial recognition AI, for example, looks at many pictures of faces and extracts mathematical relationships between various points on each face it reviews.[74] The following graphic illustrates this process:[75]



The mathematical representations labeled F7 and F8 on this graphic are stored in the system's artificial neurons.[76] As Samuelson suggests, those mathematical representations probably are not copyrightable, even though they are entirely machine generated.[77] The mathematical representations are more like facts or ideas rather than expressions. But the original image *is* reproduced, at least temporarily, to generate the mathematical representations.[78] As Ben Sobel notes, this initial reproduction of the original image is a prima facie violation of the reproduction right.[79]

Notwithstanding his acknowledgment that most AI training involves reproduction, Sobel suggests that in some cases, training data might involve only non-infringing de minimus copying.[80] As an example, Sobel argues that a human facial recognition program trained only on specific portions of human portraits might not entail reproductions of the underlying portraits.[81] Perhaps Sobel is correct in some sense. Copying is a fact-specific inquiry. But the example Sobel offers shows why de minimus

[74] *See* WILLIAM CRUMPLER & JAMES A. LEWIS, CTR. FOR STRATEGIC & INT'L STUD., HOW DOES FACIAL RECOGNITION WORK?: A PRIMER 3-6 (June 2021).

[75] *Id.* at 6 (citing Yaniv Taigman et al., Meta, Deepface: Closing the Gap to Human-Level Performance in Face Verification 4 (June 24, 2014), https://research.fb.com/publications/deepface-closing-the-gap-to-human-level-performance-in-face-verification/).

[76] *Id.*

[77] For issues relating to whether a machine can be a copyright "author," see Daniel J. Gervais, *The Machine as Author*, 105 IOWA L. REV. 2053 (2020).

[78] *See id.*

[79] Sobel, *supra* note 44, at 67. In an image-recognition system's training process, the original image's reproduction might be transitory. Once the process of decomposition and extraction begins, there is no reason for the system to retain the original image. *See infra* Section III.B.

[80] Sobel, *supra* note 44, at 67-68.

[81] *Id.* at 68.

copying is unlikely a good defense in most cases. Sobel's example is Labeled Faces in the Wild ("LFW"), a data set used for testing facial recognition applications.[82] The University of Massachusetts Amherst maintains LFW.[83]

Sobel suggests that "little copyrightable content remains in the dataset" because the dataset reproduces "only the portions of the photographs that show the subjects' faces."[84] A review of the dataset, however, shows that, absent fair use, it undoubtedly violates the copyright owners' reproduction and adaptation rights in the underlying photographs.

LFW's base images were culled from a larger set of images, called "Names and Faces," extracted by other academic facial recognition technology researchers from the commercial *Yahoo News* website.[85] Those researchers obtained their "very large data set" from "half a million news pictures" using a face detection tool.[86] A technical paper describing Names and Faces shows how the cropped photos connect to the underlying photos (which are available through links in Names and Faces):[87]

[82] *Id.* at 67.

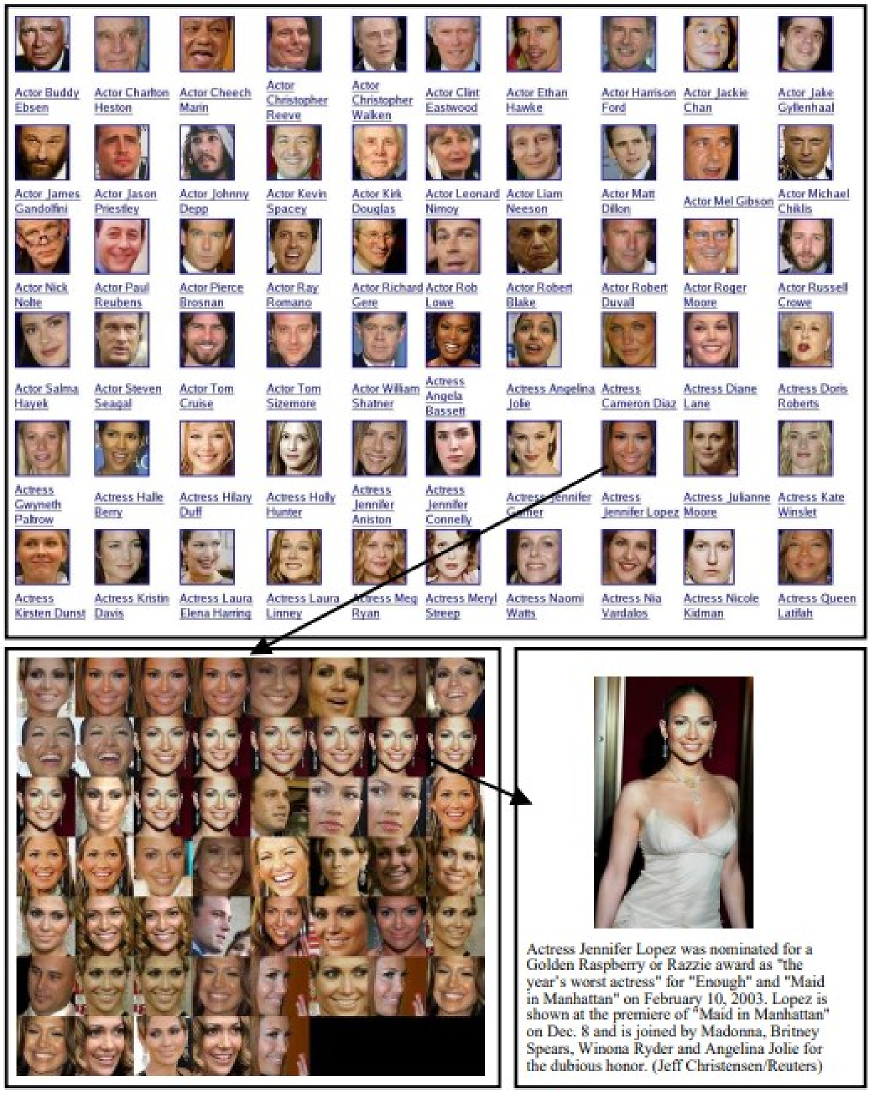
[83] *See Labeled Faces in the Wild*, UNIV. OF MASS. AMHERST, http://vis-www.cs.umass.edu/lfw/ (last visited Feb. 15, 2024).

[84] Sobel, *supra* note 44, at 67.

[85] Gary B. Huang et al., Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments 9 (n.d.), http://vis-www.cs.umass.edu/lfw/lfw.pdf.

[86] Tamara L. Berg et al., Names and Faces 5-6 (n.d.), http://tamaraberg.com/papers/journal_berg.pdf.

[87] *Id.* at 22. The LFW database can be explored and downloaded at http://vis-www.cs.umass.edu/lfw/#explore. As an example, consider the entries for Britney Spears. *Labeled Faces in the Wild: Images for Brittany Spears*, UNIV. OF MASS. AMHERST, http://vis-www.cs.umass.edu/lfw/person/Britney_Spears.html (last visited Feb. 15, 2024).



Actress Jennifer Lopez was nominated for a Golden Raspberry or Razzie award as "the year's worst actress" for "Enough" and "Maid in Manhattan" on February 10, 2003. Lopez is shown at the premiere of "Maid in Manhattan" on Dec. 8 and is joined by Madonna, Britney Spears, Winona Ryder and Angelina Jolie for the dubious honor. (Jeff Christensen/Reuters)

Contrary to Sobel's suggestion, it seems highly unlikely that a court would find that the cropped versions of these photos are noninfringing. The cropped photos still reflect the photographers' decisions about timing, pose, lighting, expression, and effects that allow copyright in the photographs.[88] These unlicensed reproductions are prima facie violations of the authors' reproduction and adaptation rights.

In addition to supporting the claim that AI training data almost always involves copying, the LFW example also shows why copyright issues relating to AI training data involve much more than the mathematical tokens generated by and stored in any one AI system. There are markets for AI training datasets, like LFW, that persist over time.[89] These training databases retain the original materials used for training. Not only is the proprietor of an AI system using the training data to create reproductions and adaptations, but so is the proprietor of the training database. This might be motivated by academic research purposes, as in the case of the LFW, or for commercial purposes by entities that license their databases for training.[90]

This Article will further explain these issues when discussing fair use.[91] But first, it discusses two reasons why this prima facie infringement of the reproduction right might not produce liability: transitory reproduction and consent.

B. Transitory Reproduction

As discussed, AI training databases gleaned from crawling and scraping usually retain copies or adaptations of the original training data that, absent consent or fair use, likely infringes the copyrights of those original data.[92]

[88] *See* Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53, 60 (1884). It is possible, though not likely, that substantial cropping could render a use de minimus and not infringing. *Cf.* Hirsch v. CBS Broadcasting Inc., No. 17 CIV. 1860 (PAE), 2017 WL 3393845, at *3 (S.D.N.Y. Aug. 4, 2017) (finding copying of a photograph in a news story was not de minimus) ("[E]ven though a fair amount of the Photo is cropped out, the average lay observer would recognize it as a copy."). In recent years, courts have found that cropping a photograph can constitute willful infringement and violation of 17 U.S.C. § 1202 if the cropping elides copyright-management information such as watermarks. *See, e.g.*, Phillips v. TraxNYC Corp., No. 21-CV-528 (LDH)(MMH), 2023 WL 1987206 (E.D.N.Y. Feb. 14, 2023); Stokes v. MilkChocolateNYC LLC, No. 22 Civ. 6786 (PAE) (RWL), 2023 WL 4447073 (S.D.N.Y. July 11, 2023). Of course, one can argue that *Burrows-Giles* incorrectly extended copyright to many kinds of photographs, but this law has been well settled since 1884. It is also certainly true that, at some point, a photo crop cannot automatically infringe—consider a crop of a photo portrait that extracts only the subject's eyeballs or nostrils or that only extracts a single pixel. The crops in LFW, however, retain the entire facial pose and lighting and do not seem close to invoking the reductio ad absurdum of minute crops.

[89] *See* discussion *infra* Section IV.C.

[90] *See* discussion *infra* Section IV.C.

[91] *See* discussion *infra* Part IV.

[92] *See supra* Section II.A.

Transitory reproduction does not apply to these databases. Specific AI systems, however, store raw training data in memory only briefly and retain only uncopyrightable mathematical tokens abstracted from the training data.[93] A line of cases running from the early computing and video game eras into the early period of digital video retransmission might suggest that "transitory" copies made during training do not infringe. Transitory reproduction would not protect training databases, but it might apply to some applications that use training data.

Section 106 of the Copyright Act of 1976 ("the 1976 Act") gives a copyright owner the exclusive right "to reproduce the copyrighted work in copies."[94] The 1976 Act defines "copies" as "material objects, other than phonorecords, in which a work is fixed by any method now known or later developed, and from which the work can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device."[95] Somewhat confusingly, under the 1976 Act, copyright is also *acquired* when a work is "fixed in [a] tangible medium of expression."[96] In its definitional section, the 1976 Act states that "[a] work is 'fixed' in a tangible medium of expression when its embodiment in a copy or phonorecord, by or under the authority of the author, is sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than transitory duration."[97] Fixation, then is implicated in both the acquisition of copyright and what comprises a potentially infringing copy under the reproduction right.[98]

Courts first focused on the word "fixed" in connection with computing and digital video technologies that make temporary "cache" copies of some content and in connection with video games that produce ephemeral displays on a screen. Early cases involving 1980s arcade video games set the stage for this argument.[99] These cases involved read-only memory chips ("ROMs")

[93] *See supra* Section II.A.

[94] 17 U.S.C. § 106.

[95] *Id.* § 101.

[96] *Id.* § 102.

[97] *Id.* § 101.

[98] 17 U.S.C. § 101. Aaron Perzanowski, however, argues that "fixed" might not mean the same thing in each context. Aaron Perzanowski, Essay, *Fixing RAM Copies*, 104 NW. U. L. REV. 1067, 1088-89 (2010).

[99] *See* Midway Mfg. Co. v. Artic Int'l, Inc., 547 F. Supp. 999, 1002 (N.D. Ill. 1982), *aff'd*, 704 F.2d 1009 (7th Cir. 1983); Stern Elecs., Inc. v. Kaufmann, 669 F.2d 852, 857 (2d Cir. 1982); Williams Elecs., Inc. v. Artic Int'l, Inc., 685 F.2d 870, 874 (3d Cir. 1982).

programmed to modify or reproduce popular arcade games.[100] Defendants argued that the games were not "fixed" because the sequence of images and sounds appeared only briefly on a screen during play and could vary in multiple ways through a player's interaction with the game.[101] Courts rejected these arguments, holding that the instructions programmed onto the original games' ROMs sufficiently fixed the games' images, patterns, and sequences, notwithstanding variation from user input.[102]

The next phase of the fixation debate involved random-access memory ("RAM"). At the dawn of the personal computing era, in the pioneering and much-derided case of *MAI Systems Corp. v. Peak Computer, Inc.*, the Ninth Circuit held that "copying" of a computer program occurs whenever the program is transferred from permanent storage to temporary RAM memory.[103] The case involved a service company, Peak, that was hired to maintain and repair computers running MAI software.[104] MAI's customers were licensed to use the software, but that license did not extend to third party maintenance companies.[105] Peak argued that it never reproduced the software because it did not copy or modify any of the files on its customers' hard drives.[106] MAI argued that a copy is made every time the computer is turned on because software files are loaded into temporary RAM memory so that the program can run.[107] The court agreed with MAI.[108] This decision, which enabled software providers to control aspects of the maintenance and repair of computer systems, attracted much scholarly and policy debate.[109]

From the video game and *MAI* cases, it seemed that the fixation with "fixed" was futile. In 2008, however, the Second Circuit in *Cartoon Network LP v. CSC Holdings, Inc.* (more generally known as *Cablevision*) decision seemingly breathed new life into the question by holding that a temporary

[100] *See Midway Mfg. Co.*, 547 F. Supp. at 1002; *Stern Elecs.*, Inc. 669 F.2d at 854; *Williams Elecs., Inc.*, 685 F.2d at 872.

[101] *See Midway Mfg. Co.*, 547 F. Supp. at 1008; *Stern Elecs., Inc.*, 669 F.2d at 855; *Williams Elecs., Inc.*, 685 F.2d at 874.

[102] *See Midway Mfg. Co.*, 547 F. Supp. at 1008; *Stern Elecs.*, 669 F.2d at 856; *Williams Elecs.*, 685 F.2d at 874.

[103] 991 F.2d 511, 519 (9th Cir. 1993).

[104] *Id.* at 517.

[105] *Id.*

[106] *Id.*

[107] *Id.* at 518.

[108] *Id.*

[109] *See* Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful*, 53 U.C. DAVIS L. REV. 893, 928-29 (2019).

"buffer" copy of a video transmission was not copyright infringement.[110] This case involved early versions of cloud-based digital video recorders ("DVRs") in a time just before the streaming revolution.[111]

One of the purposes of the 1976 Act was to harmonize U.S. copyright law with the Berne Convention's international standards concerning copyright formalities and terms.[112] Another purpose was to accommodate the new and growing cable television business.[113] Cablevision was authorized to retransmit television signals to its cable television subscribers. *Sony Corp. v. Universal City Studios, Inc.* established that under the fair use doctrine, individuals could record television programs on home video cassette recorders ("VCRs") for the purpose of "time shifting."[114] Cablevision's cloud-based DVR took a stream of its broadcast data into buffer memory, which held the data for no more than 1.2 seconds.[115] This cloud-based DVR facilitated real-time rewind for users.[116] If a customer wanted to record a program for later viewing, the data was stored in server space allocated to that customer.[117]

The Second Circuit held that the 1976 Act's definition of "fixed"

> imposes two distinct but related requirements: the work must be embodied in a medium, i.e., placed in a medium such that it can be perceived, reproduced, etc., from that medium (the "embodiment requirement"), and it must remain thus embodied "for a period of more than transitory duration" (the "duration requirement").[118]

The court distinguished *MAI* by noting that the duration requirement had not been fully litigated.[119] The court concluded that the buffer copies in

[110] 536 F.3d 121, 130 (2d Cir. 2008).

[111] *Id.*

[112] *See Copyright Timeline 1950–1997, Highlight: Congress Passes the Current Copyright Act*, U.S. COPYRIGHT OFF., https://www.copyright.gov/timeline/timeline_1950-1997.html (last visited Feb. 25, 2024).

[113] *See* REG. OF COPYRIGHTS, THE CABLE AND SATELLITE CARRIER COMPULSORY LICENSES: AN OVERVIEW AND ANALYSIS 51-52 (Mar. 1992), https://copyright.gov/reports/cable-sat-licenses1992.pdf.

[114] 464 U.S. 417, 456 (1984).

[115] *Cartoon Network LP*, 536 F.3d at 129-30.

[116] *Id.* at 135.

[117] *Id.* at 124.

[118] *Id.* at 127.

[119] *Id.* at 128.

Cablevision's cloud-based DVR were not fixed under the duration requirement and therefore could not be infringing "copies."[120]

AI training might resemble the "transitory" copies of the *Cablevision* cloud-based DVR because once an AI is trained on a dataset, the underlying data does not remain within the AI system. The *Cablevision* court, however, did not set an outer bound for what "transitory" means. In raw, unstructured AI training, any individual artifact may be ingested, deconstructed, and compared relatively quickly—maybe even comparable to the 1.2 seconds of the Cablevision ingest buffer. Best practices for AI training, however, require more time with the data because a human being is in the loop, applying metadata and adjusting the algorithms before and during the training.[121] In fact, there is now a rapidly growing industry in data annotation.[122]

The variability involved with the term "transitory" highlights a significant problem with *Cablevision*'s view of copying: the meaning of "transitory" is essentially arbitrary and infinitely malleable depending on technology and circumstances. The 1.2 seconds that Cablevision's ingest buffer used is rapid compared to unaided human capabilities, but it already seems ponderously slow compared to current data transmission and computer processing speeds. What is "transitory" to the human eye is leisurely to a powerful computer.

[120] *Id.* at 130. In *American Broadcasting Co. v. Aereo, Inc.*, 573 U.S. 431 (2014), a similar question reached the Supreme Court. The Court concluded that Aereo's cloud-based DVR system, which was comprised of an array of antennas that captured digital broadcast television, violated the "Transmit Clause" because Aereo was not a cable television provider. *Id.* at 450-51. This distinction means that *Aereo* did not address the same questions as the Second Circuit did in *Cartoon Network LP*. Cartoon Network LP v. CSC Holdings, Inc., 536 F.3d 121, 126 (2d Cir. 2008). In any event, the rapid growth of streaming services such as Netflix has moved the market from consumer-directed recording of cable or broadcast television to on-demand streaming of content hosted by streaming services. *See Cloud DVR Market by Platform, Type, and Geography – Forecast and Analysis 2023-2027*, TECHNAVIO (Nov. 2022), https://www.technavio.com/report/cloud-dvr-market-industry-analysis (webpage summarizing the report) ("The high adoption of free online video streaming is a major challenge to the global cloud digital video recording (DVR) market growth.").

[121] *See, e.g.*, Vickram Singh Bisen, *Why Data Annotation Is Important for Machine Learning and AI?*, MEDIUM (Dec. 21, 2019), https://medium.com/vsinghbisen/why-data-annotation-is-important-for-machine-learning-and-ai-5e647637c621; *Data Annotation Tools for Machine Learning (Evolving Guide)*, CLOUD FACTORY, https://www.cloudfactory.com/data-annotation-tool-guide (last visited Feb. 15, 2024).

[122] *See, e.g.*, *Top Market Reports: Data Annotation and Labeling Market – Global Forecast to 2027*, MARKETS & MARKETS, https://www.marketsandmarkets.com/Market-Reports/data-annotation-and-labelling-market-20349022.html [https://perma.cc/RG8T-MU5X] (last visited Mar. 20, 2024) (screenshot of sample page from full report) ("The global data annotation and labeling market is expected worth [sic] USD 3.6 billion by 2027, growing at a CAGR of 33.2% during the forecast period.").

What takes hours of computing time today will take seconds in a few years. When quantum computing takes hold, everything we do today will seem sloth-like. It seems that *Cablevision's* view of copying was actually shorthand for fair use and fair use is where an analysis of short-term copying for data processing purposes belongs.

In addition to these factual and doctrinal problems,[123] from the perspective of AI policy and ethics, we do *not* want proprietors of AI systems to destroy their training data.[124] If the AI is producing undesirable results, the training data might help us understand why. Further, privacy law in many jurisdictions require that data subjects whose PII was used in training data have access to that data and rights of portability and rectification.[125] At the very least, an AI system proprietor should be able to explain what, if any, of a data subject's PII was used and subsequently deleted. Transitory reproduction, then, seems a bad fit for avoiding copyright in AI training data both as a practical and policy matter, even if *Cablevision's* articulation of the doctrine in relation to cloud-based DVR technology otherwise makes sense.

C. Consent

Perhaps the most obvious and overlooked response to copyright in AI training data gleaned from the Internet is consent. Copyright owners, of course, can "authorize" others to use their works through assignments and licenses.[126] Indeed, this is how people in creative industries typically make money from copyrights.[127] Most copyright-protected content, however, is not directly monetized. Many commercial websites do not directly make money from their content but instead direct users to their products and services.[128] Such sites typically link to a terms of service that allow users to view the content through their Internet browsers but not to otherwise distribute or make copies of the content.[129] Most people who contribute online content

[123] *See supra* Section II.A.

[124] *See, e.g.*, *Blueprint for an AI Bill of Rights*, *supra* note 19, at 5 (noting that part of protection against algorithmic discrimination involves "use of representative data").

[125] *Cf.* General Data Protection Regulation, art. 16, 2016 O.J. (L 119).

[126] 17 U.S.C. §§ 106, 201(d).

[127] *See* Rich Stim, *Copyright Ownership: Who Owns What?*, STAN. LIBRS., https://fairuse.stanford.edu/overview/faqs/copyright-ownership/ (last visited Feb. 15, 2024).

[128] Kasey Kaplan, *Why Every Business Needs a Website*, FORBES (Feb. 3, 2020, 7:00 AM), https://www.forbes.com/sites/theyec/2020/02/03/why-every-business-needs-a-website/?sh=69d52d1f6e75.

[129] *See, e.g.*, *FCA Website Terms of Use*, JEEP, https://www.jeep.com/crossbrand_us/terms-of-use (last updated Apr. 6, 2021) (section 3 "License Grant," section 4 "Use Restrictions," and section 5 "FCA's Intellectual Property").

through social media sites and the like do not make money from their content; rather, they receive other social rewards in return for the nonexclusive licenses they give to hosting sites to publish their content.[130]

In both the typical commercial and social media cases, licenses are usually limited to the intended use of making the content available for others to view online—a limitation that precludes other uses, including web crawling and data scraping. But these sites are routinely crawled for the purpose of Internet searches without allegations of copyright infringement. In fact, web crawling is the foundation for Internet search engines, including Google, which indexes the web through crawling.[131] Google, unsurprisingly, never asked for anyone's permission before launching its indexing and search technology. So why are Google, Microsoft, and other Internet-search providers not liable for billions upon billions of copyright infringements?

No one knows because there has never been serious test litigation over standard web searches. There are, however, some well-known early cases involving some aspects of image searches, particularly the *Perfect 10 v. Amazon* and *Kelly v. Arriba Soft Corp.* cases.[132] These cases, discussed below, are widely considered to have established that a search is fair use.[133] But these relatively early cases, decided by just a few circuit courts, only examined specific kinds of rough image-based search capabilities. These cases seem a rickety support for the massive, globally important search business.

The more prosaic explanation is consent. In addition to their terms of service, websites conventionally include a "robots.txt" file that specifies the rules for web crawlers.[134] Most web content producers *want* their sites indexed by search engines such as Google, so there is no reason to configure the robots.txt file to the contrary or to deploy other technological protection measures—much less to sue Google for copyright infringement.

[130] *See, e.g.*, *Terms of Service: Your Content and Conduct*, YOUTUBE (Dec. 15, 2023), https://www.youtube.com/t/terms#27dc3bf5d9 (paragraph titled "License to YouTube").

[131] *See How Google Search Organizes Information*, GOOGLE SEARCH, https://www.google.com/search/howsearchworks/how-search-works/organizing-information/ (last visited Feb. 15, 2024) ("Most of our Search index is built through the work of software known as crawlers.").

[132] Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1160-63 (9th Cir. 2007); Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003).

[133] *See infra* Section II.D.

[134] *See* Alexander S. Gillis, *Web Crawler*, TECHTARGET, https://www.techtarget.com/whatis/definition/crawler (last updated Sept. 2022); *Introduction to robots.txt*, GOOGLE SEARCH CENT., https://developers.google.com/search/docs/crawling-indexing/robots/intro (last visited Feb. 15, 2024).

Some courts have held that a robots.txt file is a technological measure under the Digital Millennium Copyright Act such that circumventing the file's restrictions is unlawful.[135] Perhaps, as a few courts have held, configuring the robots.txt file so that it allows crawling is a form of express, or at least implied, license to reproduce the content to the extent necessary for the allowed purpose, such as web indexing.[136]

Perhaps, also, the express or implied consent to web crawling for search extends to crawling and scraping for AI training. This seems to motivate some of the copyleft sentiment that copyright should not restrict the use of public web content for AI training.[137]

Internet search fostered a set of norms about types of web crawling that facilitated searches and that no one wants to test. But copyright litigation over web-scraped AI training datasets, which either conform to robots.txt permissions or circumvent them, might test the current look-the-other-way ethos of crawling and data mining beyond its breaking point.[138] Indeed, this

[135] *See* Healthcare Advocates, Inc. v. Harding, Early, Follmer & Frailey, 497 F. Supp. 2d 627, 643 (E.D. Pa. 2007).

[136] *See* Field v. Google Inc., 412 F. Supp. 2d 1106, 1116 (D. Nev. 2006) (holding failure to configure metatags to prevent indexing constituted implied license for web crawler search indexing); Parker v. Yahoo!, Inc., No. 07-2757, 2008 WL 4410095 (E.D. Pa. Sept. 25, 2008) (holding failure to configure robots.txt file or to send a DMCA take-down notice constitutes implied license for web crawler indexing by web search engine). *Contra* Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537, 563-66 (S.D.N.Y. 2013) (holding that failure to configure a robots.txt file to prevent crawlers was not an implied license for a news clipping service to crawl and scrape AP's web content); Tamburo v. Dworkin, 974 F. Supp. 2d 1199, 1216 (N.D. Ill. 2013) (following *Meltwater*). Other cases reach similar results even for data not subject to copyright protection. For example, hiQ, a "people analytics" company that scraped public LinkedIn user profiles, breached LinkedIn's User Agreement, including by circumventing the robots.txt file's limitations. hiQ Labs, Inc. v. LinkedIn Corp., 639 F. Supp. 3d 944, 954-55 (N.D. Cal. 2022).

[137] "Copyleft" can be described as "a general method for making a program (or other work) free (in the sense of freedom, not "zero price"), and requiring all modified and extended versions of the program to be free as well." *What Is Copyleft?*, GNU OPERATING SYS., https://www.gnu.org/licenses/copyleft.en.html (last updated Jan. 2, 2022). Copyleft can also be used colloquially to refer to a normative sense that copyright unduly restricts access to culture and relates to the "free culture" movement. Ben Sobel describes copyleft / free-culture advocates who argue against copyright protections for AI training data as "decelerationists." *See* Ben Sobel, *Don't Give AI Free Access to Work Denied to Humans, Argues a Legal Scholar*, THE ECONOMIST (Feb. 16, 2024), https://www.economist.com/by-invitation/2024/02/16/dont-give-ai-free-access-to-work-denied-to-humans-argues-a-legal-scholar.

[138] In addition to copyright claims, crawling and scraping raises issues under the Computer Fraud and Abuse Act and under the common law of contracts, torts, property, and

seems to already be happening as the growing litigation and regulatory activity around copyright in AI training data shows. An argument grounded in implicit consent seems unlikely to prevail, certainly on a prospective basis if a copyright proprietor explicitly restricts use of its content for AI training. Fair use would be a much more secure ground apart from consent—or, as the discussion in Part IV argues, a fair use analysis suggests that the best solution is a more robust focus on consent through a combination of voluntary and compulsory licensing explicitly linked to AI training uses.

IV. Fair Use: So-Called Non-Expressive Uses

Some AI advocates argue for a broad fair use principle that would make copyrighted material generally available for AI training. These arguments mirror broader concerns about the information and research commons.[139] Such concerns are understandable, but they rest on uncertain doctrinal grounds and overlook the dynamics of AI training and application markets.

A. Non-Expressive Use: Not Quite a Doctrine

The doctrinal core of this fair use argument is non-expressive use.[140] For example, OpenAI, the creator of ChatGPT and DALL-E, argues that its use of training data is transformative because "[w]orks in training corpora were meant primarily for human consumption for their standalone entertainment value" and because the outputs of the LLMs are different than the training data.[141]

The 1976 Act lists four factors for determining whether a use is fair:

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;

privacy. *See* Benjamin L.W. Sobel, *A New Common Law of Web Scraping*, 25 LEWIS & CLARK L. REV. 147, 148 (2021).

[139] *See, e.g.*, Carroll, *supra* note 109, at 938.

[140] *See* Cullen O'Keefe et al., U.S. Pat. & Trademark Off., Dep't of Com., Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation 5 n.18 (n.d.), https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf (citing Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. 1607, 1638 (2009)); Sobel, *supra* note 44, at 51-57. Sobel notes that even if there is a non-expressive use doctrine under fair use, many existing AI applications produce expressive outputs, so the use is not really non-expressive in any event. *Id.* at 72.

[141] O'Keefe et al., *supra* note 140, at 5.

> (2) the nature of the copyrighted work;
>
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
>
> (4) the effect of the use upon the potential market for or value of the copyrighted work.[142]

Under the "purpose and character of the use" factor, according to the Supreme Court in *Campbell*, the question is "whether the new work merely 'supersede[s] the objects' of the original creation, or instead adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message; it asks, in other words, whether and to what extent the new work is 'transformative.'"[143]

Non-expressive use focuses on the way copyrighted works can function as inputs in the production of outputs that are not themselves infringing on the input works. Our simplified image recognition AI of the beach scene is a good example.[144] The training inputs include images of beaches. The AI system's output is not an image at all: it is a decision upon evaluating another image. The decision—"yes, that is a beach" or "no, that is not a beach"—does not infringe on the beach training images.

The non-expressive use concept has some intuitive appeal. A photographer cares about her beach photograph and the market for that photograph, not about the decision whether another photograph is also a beach scene. But the theoretical and practical basis for this supposed doctrine and for its application to AI training data seems shaky. At the very least, this basis cannot serve as blanket permission to exploit copyrighted works for AI training in all circumstances.

1. Book Scanning, Search Engines, and Digital Archives

The most persuasive argument for non-expressive fair use is derived from the Second Circuit's decisions in *Authors Guild v. Google, Inc.* ("*Google Books*") and *Authors Guild, Inc. v. HathiTrust*.[145] These cases involved scanning large volumes of books from academic and other libraries. *HathiTrust* involved books scanned by Google, the Internet Archive, and

[142] 17 U.S.C. § 107; *see also* O'Keefe et al., *supra* note 140, at 5.

[143] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994) (alteration in original) (citations omitted).

[144] *See supra* note 45 and accompanying photograph.

[145] Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015); Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).

Microsoft.[146] The HathiTrust website made text-to-speech versions of copyrighted books available to individuals with visual disabilities and allowed anyone to search the text of scanned books.[147] *Google Books* addressed Google's "snippet view," which allowed a user to search the full text of a copied book and returned the search term in context with a small portion of the text.[148]

In *Google Books*, the Second Circuit noted that "[c]omplete unchanged copying has repeatedly been found justified as fair use when the copying was reasonably appropriate to achieve the copier's transformative purpose and was done in such a manner that it did not offer a competing substitute for the original."[149] The "snippet view" did not allow users to piece together an entire book.[150] The court concluded that "Google has constructed the snippet feature in a manner that substantially protects against its serving as an effectively competing substitute for Plaintiffs' books."[151] The Second Circuit reached a similar conclusion about the "search" function in *HathiTrust*, along with reproductions made accessible to individuals with visual disabilities.[152]

The Google Books project can be seen as a kind of early rehearsal for today's quantitatively much larger and qualitatively much more disruptive arguments about AI training data. In 2009, Professor Matthew Sag first observed that some uses of expressive inputs for purposes significantly beyond their original expressive purpose could constitute non-expressive fair use.[153] Similarly, in his paper, *Copyright for Literate Robots*, Professor James Grimmelmann argues that the Second Circuit's *Google Books* decision supports the argument that "[b]ulk nonexpressive uses," including "bulk reading" by machines, "are fair uses."[154] In his paper, Grimmelmann sketches what he believes existing doctrine says, not necessarily what it should say.[155] He acknowledges that "[i]t is easy to see how bulk nonexpressive copying promotes progress in artificial intelligence," but this, he says, "arguably increases the chances that humanity will meet a sudden, violent, and

[146] *See HathiTrust*, 755 F.3d at 90.

[147] *Id.* at 91.

[148] Authors Guild v. Google, Inc., 804 F.3d at 221.

[149] *Id.*

[150] *Id.* at 221-22.

[151] *Id.* at 222.

[152] *See id.* at 221 (citing *HathiTrust*, 755 F.3d at 98).

[153] Sag, *supra* note 140, at 1638.

[154] James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 666-67 (2016).

[155] *Id.* at 657, 674, 681.

extremely unpleasant end."[156] Professors Mark Lemley and Bryan Casey make a similar argument, although seemingly with less hesitation about the dangers of AI, in their paper *Fair Learning*.[157]

It is not so clear, however, whether existing doctrine says anything so broad about "bulk non-expressive uses." The Second Circuit's focus in *Google Books* and *HathiTrust* was on the market for the copyrighted work, not on the degree of expression in the allegedly infringing use.[158] For the Second Circuit, the "amount and substantiality of the portion used" factor must be read in tandem with the effect on the market factor.[159] The court credited Google's and HathiTrust's *factual* arguments that search snippets, which the full-text scans enabled, would not erode the market for complete published books.[160] This is not a sweeping doctrinal conclusion about other kinds of "bulk non-expressive uses," much less about AI or robot uses.

Lemley and Casey also argue that the bulk, non-expressive use exception is rooted in early Internet search engine cases.[161] They claim the "non-expressive use" exception is "the reason most automated search and analysis tools exist in the first place."[162] The cases, however, are not so clear.

In *Perfect 10 v. Amazon*, an early image-based search case that is often relied upon to support a non-expressive use exception, the Ninth Circuit held that Google did not infringe the display nor distribution rights by providing hyperlinks to full-sized photos housed on Perfect 10's servers but that thumbnail versions of the images could infringe these rights.[163] This view of the scope of the display and distribution rights was called the "server test."[164] However, the court held that Google's use of the images was fair use.[165] Under the purpose and character of the use factor, the court concluded that

[156] *Id.* at 678.

[157] Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 745, 748, 769 (2021).

[158] Authors Guild v. Google, Inc., 804 F.3d 202, 221 (2d Cir. 2015); Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 95, 99 (2d Cir. 2014).

[159] *See* Authors Guild v. Google, Inc., 804 F.3d at 218-21.

[160] *See id.* at 222.

[161] Lemley & Casey, *supra* note 157, at 762.

[162] *Id.* "Most" here seems a substantial overstatement. For example, some important automated search and analysis tools, such as Westlaw and Lexis, index material that is mostly in the public domain. What Lemley and Casey seem to reference are tools that mine data from the public Internet, such as Google's search engine. They also identify the Digital Millennium Copyright Act's protections against secondary liability as a key driver. *Id.*

[163] Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1160-61 (9th Cir. 2007).

[164] *Id.* at 1159-60.

[165] *Id.* at 1160.

"[a]lthough 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."[166]

The court's reasoning in *Perfect 10* seems similar to the case for fair use of AI training data: the training process transforms the image into a mathematical cipher for decision making. The claims in *Perfect 10*, however, only involved the display and distribution rights, not the rights of reproduction or derivative work.[167] This is an important distinction because of how the technology at issue in *Perfect 10* worked. Training data, in contrast to the search process, is not displayed or distributed as a pointer to other information. It is copied wholesale to make what is arguably a derivative work.

Another early search engine case, *Kelly v. Arriba Soft Corp.*, is closer to the point.[168] Arriba Soft crawled and scraped images to generate low-resolution image thumbnails that were used in its search engine.[169] Under the first fair use factor, the Ninth Circuit stated that "Arriba's search engine functions as a tool to help index and improve access to images on the internet and their related web sites" and that the thumbnails were not useful for artistic purposes because of their low resolution.[170] This functional difference, along with the lack of a market for the use of Kelly's images in search engines, led the court to find the use was fair use.

Proponents of non-expressive fair use argue that early search engine cases involve expressive inputs used for significantly different purposes than their original expressive purpose.[171] The functional change from aesthetically pleasing work to mathematical cipher in a search engine could be similar to the functional change when information is used to train an AI, at least involving things like photographs. On the other hand, because some of the copyrighted works used to train an AI are *meant for training* human beings, the transformation in purpose is not so dramatic.

Some advocates of a broad non-expressive fair use rule further argue that the rule is necessary because copyright is not designed to capture, and nor should it capture, all the potential positive spillovers of copyrightable

[166] *Id.* at 1165.

[167] *Id.* at 1159.

[168] *See* 336 F.3d 811, 815 (9th Cir. 2003).

[169] *Id.* at 815-16.

[170] *Id.* at 818.

[171] *See, e.g.*, Sag, *supra* note 140, at 1616-17.

expression.[172] Under this view, for example, the author of a novel can capture the value conferred on a reader who publishes the novel for the purposes of her own entertainment, education, and enlightenment, but the author is not entitled to the spillover value of the text as an input for an Internet search engine or as a book search "snippet." The same logic, they argue, applies to the use of the novel in AI training data.[173]

This comparison, however, is too quick. An Internet search engine and a book search snippet allow end users to learn factual information about the novel, including how to purchase the text or borrow it from a library that has purchased it. Some AIs may use training data only to produce basic factual information, analogous to a search engine, or to make basic decisions unrelated to the author's original purposes relating to her readers and potential readers. Other AIs, including LLMs such as OpenAI and DALL-E, may use training data to produce texts, images, and other outputs that serve the same or similar purposes as the copyrighted work. These applications seem much more like derivate works—which by statute are expressly within the range of spillovers the author can and should be able to capture—rather than more distant, unrelated applications that might fall outside that range.[174]

Moreover, in both *Perfect 10* and *Kelly*, the transformativeness factor was closely tied to the effect on the market factor. The more transformative the use, the less likely the use falls within the zone of the input work's existing or reasonably possible markets.[175] At the time of those cases, there was no market for the licensing of copyrighted works for search engine listings. The early search engine cases, then, could fit Wendy Gordon's paradigm for fair use as a response to market failure.[176] The question whether the expressive works were used for different, non-expressive purposes was intimately tied to whether there was a viable market for the input works as inputs for such purposes. There is now a growing market for use of existing and new works of all kinds as AI training data, which suggests that the early search engine cases may not apply.[177] In any event, the early search engine cases focus on

[172] *See* Lemley & Casey, *supra* note 157, at 749; Brett M. Frischmann & Mark A. Lemley, Essay, *Spillovers*, 107 COLUM. L. REV. 257, 258 (2007).

[173] Lemley & Casey, *supra* note 157, at 762-63.

[174] *See* 17 U.S.C. § 106.

[175] In some ways, this resembles a cross-elasticity of demand analysis for purposes of market definition in antitrust law.

[176] Wendy J. Gordon, *Fair Use as Market Failure: A Structural and Economic Analysis of the* Betamax *Case and Its Predecessors*, 82 COLUM. L. REV. 1600, 1601-02 (1982).

[177] *See* Sobel, *supra* note 44, at 53, 80-81.

specific uses and markets and do not announce a generally applicable rule of non-expressive fair use.[178]

2. Other Bases for Non-Expressive Use

In addition to the early search engine cases, *Google Books*, and *HathiTrust*, Matthew Sag and other scholars posit a concept of non-expressive fair use based on various other snippets of copyright doctrine, including the collective work right as discussed in the Supreme Court's opinion in *New York Times Co. v. Tasini*.[179] *Tasini* dates to an era when

[178] Further, it is notable that in each of the earlier circumstances regarding Internet searches and book snippets, the technology giant Google mediated the positive spillovers to end users. In doing so, of course, Google took its own significant share of those spillovers. The outcome of those cases may reflect the value Google added by developing the technology needed to facilitate Internet searches and book snippets, but they may also reflect or be distorted by Google's market dominance. *See* discussion *infra* Section IV.C.

[179] Sag, *supra* note 140, at 1631; New York Times Co. v. Tasini, 533 U.S. 483 (2001). Sag also focuses on the idea-expression dichotomy, the collective work right, the substantial similarity test for infringement, and the rejection of intermediate copying claims in the entertainment industry to demonstrate that the right of reproduction protects only expressive substitution (i.e., reproduction that is available to the public). Sag, *supra* note 140, at 1628-36. Concerning early draft scripts or song versions in cases involving plays, movies, and music, courts sometimes note that infringement is based only on a film as broadcast so that preliminary scripts do not matter or that a defendant might avoid infringement by making changes before the work is broadcast. *Id.* at 1636 (citing Davis v. United Artists, Inc., 547 F. Supp. 722, 724 n.9 (S.D.N.Y. 1982); Warner Bros., Inc. v. Am. Broad. Cos., 720 F.2d 231, 241 (2d Cir. 1983); Madrid v. Chronicle Books, 209 F. Supp. 2d 1227, 1234 (D. Wyo. 2002); Walker v. Time Life Films, Inc., 615 F. Supp. 430, 434 (S.D.N.Y. 1985); Eden Toys, Inc. v. Marshall Field & Co., 675 F.2d 498, 501 (2d Cir. 1982); Durham Indus., Inc. v. Tomy Corp., 630 F.2d 905, 913 & n.11 (2d Cir. 1980)). In such cases, however, the issue was not wholesale copying of the underlying work or even any literal reproduction at all. The issue in these cases was whether nonliteral similarities in things like scene structure, sequence of events, and characters added up to unlawful copying. *See, e.g.*, Huie v. Nat'l Broad. Co., 184 F. Supp. 198, 200 (S.D.N.Y. 1960) (refusing to consider intermediate scripts) ("We can put aside the question of slavish copying because there is no suggestion of it here."). Courts usually do not allow the plaintiff to introduce comparisons based on "lists of random similarities and on earlier scripts of the screenplay" because such evidence is usually considered an unreliable measure of any nonliteral similarities in the work alleged to infringe. *See, e.g.*, *Walker*, 615 F. Supp. at 434. In other words, the plaintiffs' claims in these cases were not that the earlier scripts and the like themselves infringed but that the earlier versions provided some evidence of why the *final* version infringed. This is an interesting and knotty evidentiary question, but it falls far short of supporting a publication requirement for the reproduction or derivative rights. The substantial similarity for infringement likewise does not address the question of supposedly non-expressive uses. As Sag notes, the basic test for substantial similarity in cases

newspapers such as the *New York Times* were just beginning to digitize their past and current print editions.[180] Text-only digital copies were made available on commercial databases such as NEXIS and on CD-ROMs.[181] Previously, newspapers were archived on the analog media of microfilm or microfiche, copies of which libraries could obtain and index.[182]

In *Tasini*, the plaintiffs were independent journalists who had contributed articles to publications like the *New York Times*.[183] They argued that their publication agreements only permitted the use of their copyrighted works as part of a collective work—the print newspaper—and not as part of databases through which articles could be individually searched and viewed apart from their publication in the collective work.[184]

When *Tasini* was heard in 2001, it was viewed as a watershed moment in the developing Internet era: would collective work publishers like the *New York Times* be forced to engage in burdensome spade work to identify decades of past writers or their heirs, and pay potentially crippling new royalties to those parties, so that the public could easily search and access these documents in digital formats?[185] On the writers' side of the argument, would powerful legacy publishers such as the Times, in league with big database companies such as NEXIS, control the Internet's development and the public's ability to learn about history, or would that power disperse down

of nonliteral infringement is how the works appear to the consuming public. Sag, *supra* note 140, at 1633. This does not suggest, however, that the copyright author must make her work public to secure protection for the reproduction right. The word "public" here refers not to publication or the author's reputation but to the market for the copyrighted work. *Id.* at 1633 (citing Arnstein v. Porter, 154 F.2d 464, 473 (2d Cir. 1946) ("The [copyright owner's] legally protected interest is not, as such, his reputation . . . but his interest in the potential financial returns from his compositions . . . ."). In *Arnstein*, the court concluded that the market was the "lay public" rather than expertly trained musicians because "lay listeners . . . comprise the audience for whom such popular music is composed." *Id.* A copyright author is entitled to damages relating to both existing and potential markets. 17 U.S.C. § 504. The "effect on the market" factor in fair use analysis likewise considers both existing and potential markets. The copyright owner therefore has some right to exclude even in markets she has not yet entered. *See, e.g.*, Rogers v. Koons, 960 F.2d 301, 312 (2d Cir. 1992) (holding no fair use for sculptures based on photographs even though photographer had not yet entered sculpture market).

[180] *Tasini*, 533 U.S. at 489-90.

[181] *Id.* at 490.

[182] *Id.* at 517.

[183] *Id.* at 489.

[184] *Id.* at 486.

[185] *See* Dina Marie Pascarelli, *Electronic Rights: After* Tasini *Who Owns What, When?* Tasini v. New York Times, 8 DEPAUL J. ART, TECH. & INTELL. PROP. L. 45, 76-77 (1997).

to individual writers?[186] Iconic American historian and filmmaker Ken Burns even weighed in with an amicus brief.[187]

The Court held that an agreement to contribute a work as part of a collective work includes only the rights of reproduction and distribution as part of the collective work.[188] This, the Court said, was clear from section 201(c) of the 1976 Act, which states:

> Copyright in each separate contribution to a collective work is distinct from copyright in the collective work as a whole, and vests initially in the author of the contribution. In the absence of an express transfer of the copyright or of any rights under it, the owner of copyright in the collective work is presumed to have acquired only the privilege of reproducing and distributing the contribution as part of that particular collective work, any revision of that collective work, and any later collective work in the same series.[189]

As Sag noted, the Court distinguished between collective works and databases with reference to how the content appears to an ordinary user.[190] Sag concluded that the Court thereby "reinforced that expressive communication to the public is the touchstone of copyright infringement."[191]

Tasini, however, was a case primarily about *transfers* and only secondarily about infringement. Section 201 governs transfers of rights.[192] As Justice Stevens noted in a dissent joined by Justice Breyer, the case "raise[d] an issue

[186] *See Tasini*, 533 U.S. at 505.

[187] *Id.* The ultimate result was much less earth shattering—or rather, the earth-shattering events were something completely different. The newspapers lost and had to negotiate a settlement fund to include older content from independent journalists. The standard terms for contributor agreements changed to include other database rights along with the collective-work right. *See* Adam Liptak, *Justices Reinstate Settlement with Writers*, N.Y. TIMES (Mar. 2, 2010), https://www.nytimes.com/2010/03/03/business/media/03bizcourt.html; *see also* Eric P. Schroeder et al., *When Copyright First Met the Digital World: A Retrospective and Discussion of* New York Times v. Tasini*, 533 U.S. 483 (2001)*, COMM. LAW., Summer 2021, at 14, 28 n.3. Meanwhile, Web 2.0, with its blogs and podcasts, and the social-media revolution further disrupted every established model of journalism. Michael Karanicolas, *Disrupting Journalism: How Platforms Have Upended the News*, COLUM. JOURNALISM REV. (Feb. 13, 2023), https://www.cjr.org/special_report/disrupting-journalism-how-platforms-have-upended-the-news-intro.php.

[188] *Tasini*, 533 U.S. at 488.

[189] *Id.* (quoting 17 U.S.C. § 201(c)).

[190] Sag, *supra* note 140, at 1632; *Tasini*, 533 U.S. at 499, 501-02.

[191] Sag, *supra* note 140, at 1632.

[192] *See* 17 U.S.C. § 201.

of first impression concerning the meaning of the word 'revision' as used in § 201(c)."[193] The majority examined how the print, microfiche, and database versions appeared to the public to assess whether the database was a "revision" of a collective work or something new.[194] There was no dispute that if the agreements the authors executed did not cover the databases as "revisions" of collective works, the resulting reproduction and distribution would be infringing. Nothing in either the majority or dissenting opinions suggested a broad right of non-expressive use.

Sag also emphasized the idea-expression dichotomy in favor of non-expressive use.[195] In *Feist Publications, Inc. v. Rural Telephone Service Co.*, the Supreme Court addressed this sort of question in the context of catalog and database protection.[196] The *Feist* Court considered how to treat the decidedly old-school technology of telephone white pages.[197] As a result of Feist's reading of the idea-expression dichotomy, under U.S. law, individual facts or data points within databases are not protectible—a position at odds with the law in Europe and other parts of the world.[198]

It is true that the idea-expression dichotomy could be relevant to some infringement claims against "copyright-reliant technologies," including today's AI systems. If an AI is trained on nothing but tables of historical data (say, for example, stock prices) the idea-expression dichotomy would become important. The issue might first arise as to the copyrightability of the underlying works.[199] It might then arise under the "nature of the copyrighted work" and "amount and substantiality of the portion used" fair use factors.[200] In addition, the idea-expression dichotomy could also be relevant to a claim that the mathematical tokens *resulting from* AI training are copyrightable. But as noted, the process of creating such tokens begins with reproducing

[193] *Tasini*, 533 U.S. at 506 (Stevens, J., dissenting).

[194] *Id.* at 499.

[195] Sag, *supra* note 140, at 1631; 17 U.S.C. § 102(b).

[196] 499 U.S. 340, 343-44 (1991).

[197] *Id.*

[198] *See Directive 96/9/EC of the European Parliament and of the Council, of 11 March 1996 on the Legal Protection of Databases,* WIPO (Mar. 11, 1996), https://www.wipo.int/wipolex/en/text/126788, as anticipated in Article 5 of the WIPO Copyright Treaty of 1996. In the early 2000s, database protection bills were proposed in Congress but never gained substantial support. *See* Statement of David O. Carson, General Counsel, United States Copyright Office, on the Database and Collections of Information Misappropriation Act of 2003, at 1 (Sept. 23, 2003), https://www.copyright.gov/docs/regstat092303.html.

[199] 17 U.S.C. § 102.

[200] *See id.* § 107.

text, image, video files, and the like.[201] In most cases, this content undoubtedly passes the low threshold of "expression" under U.S. copyright law, and the entirety of the files or substantial portions of the files are ingested.[202]

3. The Digital Elephant in the Room and the Fair Use Mouse: Computer Software and APIs

The early to mid-Internet era cases previously discussed concerned aspects of digital database technologies apart from the software that makes those technologies run. Copyright protection for software is the large statutory elephant in the room looking askance at claims of non-expressive fair use.[203] Computer code is usually invisible to the public. In many contemporary software-as-a-service cloud applications, the code remains on servers that the copyright owner or its agents control.[204] Sag suggests that software is a statutory anomaly that should not dilute his broader argument.[205] But as a matter of statutory interpretation, there is no doubt that computer code is copyrightable, so the 1976 Act cannot be read to include a broad fair use protection for all non-expressive uses.[206]

The Supreme Court's recent decision in *Google LLC v. Oracle America, Inc.*, however, could signal more fair use latitude for at least some types of code inputs.[207] In that case, the Court found that Google's use of the Oracle Java Application Programming Interfaces ("APIs") was transformative because Google used the APIs "to create a new platform [i.e., Android] that could be readily used by programmers."[208] The Court noted that fair use "can play an important role" in balancing statutory copyright for software against other interests in copyright law.[209] According to the Court, fair use (at least

[201] *See supra* Section III.C.

[202] Compare the image database examples at *supra* Section III.A. *See supra* note 179.

[203] Sag, *supra* note 140, at 1638; 17 U.S.C. §§ 101, 117 (defining "computer programs" and discussing the limitation on exclusive rights for computer programs).

[204] Wesley Chai, *Software as a Service (SaaS)*, TECHTARGET, https://www.techtarget.com/searchcloudcomputing/definition/Software-as-a-Service (last visited Feb. 16, 2024).

[205] Sag, *supra* note 140, at 1638.

[206] The debate over whether computer programs were already included in the general language of section 106 of the 1976 Act before the 1980 amendments adding section 117, which specify limitations on rights in computer programs, is interesting but moot. *See* Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1198-99 (2021) (discussing history of copyright protection for computer programs).

[207] *Id.* at 1199.

[208] *Id.* at 1203.

[209] Id. at 1198.

regarding software) can "help to distinguish among technologies[,] . . . distinguish between expressive and functional features of computer code," and balance the need for incentives to create against "unrelated or illegitimate harms in other markets or to the development of other products."[210]

APIs are portions of code that allow application programs to interface with a device's operating system.[211] An operating system provides access to and control over a computing device's processing capabilities and hardware functions.[212] The proprietor of an operating system, application, or piece of hardware may make APIs available, either for free or under the terms of a license, so that other developers and consumers can create compatible applications or devices.[213]

Java was developed as a lightweight programming language for applications on devices such as television set-top boxes,[214] but it became widely used for web-based and desktop computer applications.[215] Google copied portions of some Java APIs without a license.[216] According to the Court, the portions copied included only "declaring code"—essentially a function's name—and not the "task-implementing programs" that would be called upon by the declaring code.[217] This meant that programmers familiar with Java could use well-known declaring code to implement functions in the Android operating system.[218]

[210] *Id.*

[211] *See id.* at 1191.

[212] *See id.* at 1190.

[213] *See id.* at 1190-91.

[214] Abhinandan Bhatnagar, *The Complete History of Java Programming Language*, GEEKS FOR GEEKS, https://www.geeksforgeeks.org/the-complete-history-of-java-programming-language/ (last updated Jan. 8, 2024).

[215] *See* Google v. Oracle, 141 S. Ct. at 1190.

[216] *Id.*

[217] *Id.* at 1192-94. As Joshua Bloch and Pamela Samuelson have noted, "declaring code" is a misleading phrase. Joshua Bloch & Pamela Samuelson, *Some Misconceptions About Software in the Copyright Literature*, *in* CSLAW '22: PROCEEDINGS OF THE 2022 SYMPOSIUM ON COMPUTER SCIENCE AND LAW 131, 133 (Ass'n for Computing Mach., 2022), https://dl.acm.org/doi/pdf/10.1145/3511265.3550449. It might have been preferable for the courts below, and the Supreme Court, to have recognized that "declarations" are not copyrightable and/or that Google did not copy declarations from the Java source code. *See id.* at 134-36. As discussed in Part III *supra*, however, this would not mean that AI training data is non-infringing. There is no dispute that a reproduction must be made of the AI training data in order to produce algorithmic tokens that are from an AI's "brain." Those tokens may not be copyrightable, but that is not the issue concerning training data.

[218] *See* Google v. Oracle, 141 S. Ct. at 1194.

This kind of use, the Court stated, "was consistent with that creative 'progress' that is the basic constitutional objective of copyright itself."[219] The Court also found that the amount and substantiality of the portion used was related to its purpose of permitting "programmers to make use of their knowledge and experience using the Sun Java API when they wrote new programs for smartphones with the Android platform."[220] The effect on the market, according to the Court, favored fair use because Android was unlikely to be able to compete in the operating system market and because Google's development of the Android platform benefitted the consuming public.[221]

The effect on the market analysis in *Google v. Oracle* thereby resembled the kind of consumer welfare inquiry made in first-generation antitrust cases involving computer operating systems and web browsers.[222] The Court was concerned not only with the licensing market for Java and Java APIs but also with whether restrictions on access to the APIs would limit competition in the broader operating system market.[223]

While there are some surface parallels between *Google v. Oracle's* treatment of APIs and the use of copyrighted materials as AI training data, the underlying concerns are quite different. Training data resemble APIs because both are not themselves user applications but are necessary to facilitate user applications.[224] But APIs are good for nothing other than serving as APIs.[225] AI training data—aside from synthetic data—primarily serve other functions as images, text, videos, and sounds. The use of APIs in software development is arguably non-expressive use because APIs are functional rather than expressive by design.[226] The copyrighted material used to train AI, in contrast, is by definition expressive.

Further, the developers who create the API are the same developers who create the operating systems, software, or devices to which the API

[219] *Id.* at 1203.

[220] *Id.* at 1205.

[221] *See id.* at 1208.

[222] *See* United States v. Microsoft Corp., 253 F.3d 34, 34 (D.C. Cir. 2001).

[223] *See id.* at 60.

[224] In most cases, no single piece of data is necessary to train an AI. Ryan Sevey, *How Much Data Is Needed to Train a (Good) Model?*, DATAROBOT (Aug. 4, 2017), https://www.datarobot.com/blog/how-much-data-is-needed-to-train-a-good-model/. At least for initial training, the scale of data used for training far exceeds the scale of a set of APIs even for a complex package like Java.

[225] *See What Is an API?*, RED HAT (June 2, 2022), https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces.

[226] *See id.*

interfaces.[227] If the developer can control the APIs, it can control secondary markets for systems, applications, and devices that interface with the underlying product.[228] When the underlying product is central to a technological ecosystem—like Java—restricting fair use could raise the quasi-antitrust concerns suggested by the *Google v. Oracle* majority. Such control is impossible for any individual copyright-holder in a typical AI training scenario. Training data repositories such as Common Crawl and LAION are drawn from billions of individual sources, and there is no plausible claim that any one source is necessary to develop a competitive product.

In sum, parts of copyright case law support a concept of non-expressive fair use, but it is hardly a clear or well-established concept. *Google v. Oracle* bolsters the claim that adapting a copyrighted input for a different purpose might be fair use, at least as to computer code, which is inherently close to the line of copyrightability set by the idea/expression, merger, and functionality doctrines.[229] Yet there are significant differences between APIs and the multifarious works that may be used as AI training data.

B. The Warhol Effect

Enter *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, the Court's most recent foray into fair use.[230] Although that case involves a clearly expressive use, it further complicates things for non-expressive fair use as applied to AI training data. That case involved Andy Warhol's pop art silkscreen portrait of the musician Prince based on a photograph taken by rock photographer Lynn Goldsmith.[231] The Warhol Foundation argued that Warhol's treatment of the photograph, in a style for which Warhol became famous, was transformative.[232] In an opinion written by Justice Sotomayor, the Court stated that the first fair use factor "focuses on whether an allegedly infringing use has a further purpose or different character, which is a matter of degree, and the degree of difference must be weighed against other considerations, like commercialism."[233] Although transformativeness—what

[227] *See id.*

[228] *See id.*

[229] *See* Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1209 (2021).

[230] *See* 598 U.S. 508, 516 (2023).

[231] *Id.* at 515.

[232] *Id.* at 525.

[233] *Id.*

Justice Sotomayor called "new expression"—"may be relevant . . . it is not, without more, dispositive of the first factor."[234]

Justice Sotomayor noted that the illustrative fair use purposes in section 107 of the 1976 Act offer paradigmatic examples of uses that are not merely substitutions for the underlying work.[235] But even new works that are in some sense transformative can fall within the scope of the copyright owner's right to control derivative works.[236] This is evident in the statutory definition of a derivative work, which includes "any other form in which a work may be recast, *transformed*, or adapted."[237] Therefore,

> an overbroad concept of transformative use, one that includes any further purpose, or any different character, would narrow the copyright owner's exclusive right to create derivative works. To preserve that right, the degree of transformation required to make "transformative" use of an original must go beyond that required to qualify as a derivative.[238]

Rooted in this discussion of the tension between transformative fair use and transformative derivative works, Justice Sotomayor offered two elements to consider under the first factor in addition to transformativeness: (1) whether the use is commercial; and (2) the purpose of the use.[239] If the use is commercial, this is not dispositive, but it cuts against fair use.[240] If the use has a "distinct purpose" that "furthers the goal of copyright, namely, to promote the progress of science and the arts, without diminishing the

[234] *Id.*

[235] *See id.* at 528.

[236] *See id.* at 529.

[237] *Id.* (quoting 17 U.S.C. § 101) (emphasis added).

[238] *Id.* The Court argued that "*Campbell* cannot be read to mean that § 107(1) weighs in favor of any use that adds some new expression, meaning, or message . . . . Otherwise, 'transformative use' would swallow the copyright owner's exclusive right to prepare derivative works." *Id.* at 541.

[239] *Id.* at 531.

[240] *Id.* Pamela Samuelson suggests that this part of Justice Sotomayor's opinion is dicta that may not significantly impact later fair-use cases. Pamela Samuelson, *How to Distinguish Transformative Fair Uses from Infringing Derivative Works?*, KLUWER COPYRIGHT BLOG (June 5, 2023), https://copyrightblog.kluweriplaw.com/2023/06/05/how-to-distinguish-transformative-fair-uses-from-infringing-derivative-works/ (stating that some courts and commentators "are likely to give [the *Warhol* decision] a very broad interpretation, while others may argue that it is a much narrower ruling than some dicta in Justice Sotomayor's opinion might suggest"). I offer no normative comment on whether Justice Sotomayor's opinion is best interpreted one way or the other, except to note that her language does not seem like throw-away dicta.

incentive to create," such as parody or satire, this weighs in favor of fair use.[241] In summary, Justice Sotomayor stated,

> [T]he first fair use factor considers whether the use of a copyrighted work has a further purpose or different character, which is a matter of degree, and the degree of difference must be balanced against the commercial nature of the use. If an original work and a secondary use share the same or highly similar purposes, and the secondary use is of a commercial nature, the first factor is likely to weigh against fair use, absent some other justification for copying.[242]

Thus, the inquiry into purpose is not a subjective account of the user's intent but rather "an objective inquiry into what use was made, *i.e.*, what the user does with the original work."[243]

Applied to Warhol's treatment of Goldsmith's photo, the Court determined that the original photo and Warhol's treatment of the photo served the same purpose of illustrating magazine stories about Prince.[244] This similarity of purpose paired with the commercial nature of Warhol's use weighed against fair use.[245] In response to concerns that Justice Sotomayor's seemingly narrower reading of transformativeness would stifle future creativity, she responded that "[i]t will not impoverish our world to require AWF to pay Goldsmith a fraction of the proceeds from its reuse of her copyrighted work. Recall, payments like these are incentives for artists to create original works in the first place."[246]

Warhol's limited view of transformativeness seems inconsistent with *Google v. Oracle's* more expansive view. The *Warhol* Court attempted to distinguish *Google v. Oracle* in several ways. First, the Court noted that "in applying the fair use provision, 'copyright's protection may be stronger where the copyrighted material . . . serves an artistic rather than a utilitarian function.'"[247] Because the Java code at issue in *Google v. Oracle* was "primarily functional," it was more difficult to determine the line between unlawful copying and fair use.[248] Further, Google put the Java APIs to use in

[241] *Andy Warhol Found. for the Visual Arts*, 598 U.S. at 531.

[242] *Id.* at 532-33.

[243] *Id.* at 545.

[244] *Id.* at 545-46.

[245] *Id.* at 537.

[246] *Id.* at 549.

[247] *Id.* at 527 (quoting Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1197 (2021)).

[248] *See id.* at 533 n.8.

a "distinct and different computing environment," that is, in an operating system built for mobile devices rather than in desktop and laptop computers.[249] The dissent, authored by Justice Kagan and joined by Justice Roberts, found this effort to distinguish *Google v. Oracle* unpersuasive, particularly since the *Google* Court mentioned Andy Warhol's "Campbell Soup" can graphics as paradigmatic of transformative use.[250]

Applying the *Warhol*'s additional elements to AI training data likely will not yield predictable results. Regarding the first element, some AI applications are non-commercial, but many are and will be commercial (or are and will be embedded into commercial products). Even many free AI products, including ChatGPT and DALL-E, collect user data that can be exploited by the proprietor, so the products are not actually free.[251] This factor usually will cut against fair use.

Regarding the second element, some AI applications might "further[] the goal of copyright, namely, to promote the progress of science and the arts," but others might not.[252] An AI application, such as a text or image generator, produces outputs that resemble traditional domains of copyright policy—text and images. But nonetheless, a machine generates those outputs. The 1976 Act accounts for the use of machines to fix, store, copy, distribute, and transmit copyrightable works,[253] but it does not anticipate that machines could create copyrightable expression. Courts and commentators are only just beginning to grapple with whether AI-generated content is copyrightable, but the best answer is that it is not.[254]

As noted in relation to the early search engine cases, some AI applications produce expressive content while others do not produce any creative output at all.[255] The basic function of many AI applications is to make predictions and decisions: is this an image of a beach or a desert; should the car turn left here; does Alice qualify for a mortgage; is Bob a potential candidate for this job; does this circuit board pass quality control; what advertisement will appeal to a user; and so on. Although the copyrighted inputs used for AI training were employed for reasons well beyond the purposes of their original

[249] *Id.* (quoting Google v. Oracle, 141 S. Ct. at 1203).

[250] *Id.* at 572 (Kagan, J., dissenting).

[251] *See Terms of Use*, OPENAI (Nov. 14, 2023), https://openai.com/policies/terms-of-use [https://perma.cc/DQ8H-XWLB]; *Privacy Policy*, OPENAI (Nov. 14, 2023), https://openai.com/policies/privacy-policy [https://perma.cc/XQX8-8SBP].

[252] *Andy Warhol Found. for the Visual Arts*, 598 U.S. at 510.

[253] 17 U.S.C. § 102.

[254] *See* Gervais, *supra* note 77, at 2079.

[255] *See supra* Section IV.A.

creation, the purpose of the AI training was not that of copyright—that is, the publication of more expressive content. Under *Warhol*'s reading of the first factor, then, it seems that the "non-expressive" character of the use cuts *against* fair use under these circumstances. Perhaps this is not a result the *Warhol* majority intended, but it suggests that assessing fair use for AI training inputs after *Warhol* will prove complicated.

C. The Markets in Google and Warhol and the Markets for AI Training Data

1. Transformativeness and the Effect on the Market

The *Warhol* Court, consistent with *Google v. Oracle*, *Campbell*, and other important fair use cases, recognized that the nature and character of the use factor is closely tied to the fourth factor, the effect on the market for the copyrighted work.[256] Although both *Warhol* and *Google v. Oracle* focused mostly on the first factor, it is possible to understand these cases more clearly through the fourth factor.

Similarly, fair use cases involving copyrighted AI training data might turn on whether there are existing or prospective markets for copyrighted text, images, and other content that can be repurposed as AI training data. When the present generation of text and image generators were trained, perhaps some of those markets were not yet on the horizon—but even then, there is evidence that the Google Books project anticipated later uses of scanned books for AI training.[257] It is easy to see how such markets could be plausible and beneficial since the dynamics of AI training are becoming more publicly known.

According to the *Google v. Oracle* Court, the jury could have found that there was a market for the Java APIs as a whole but not for declaring functions apart from the substantive routines called by those declarations.[258] Oracle was not in the business of using Java to create a mobile operating system.[259] The jury also could have concluded that Google's use of some declaring functions for the convenience of developers did not appreciably affect existing or prospective markets for Java.[260] Further, the jury could have found that Google's Android operating system, though it incorporated some

[256] *Andy Warhol Found. for the Visual Arts*, 598 U.S. at 536 n.12.

[257] *See* Lee et al., *supra* note 28 (manuscript at 99).

[258] *See* Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1206-07 (2021). The language is equivocal because the question posed to the jury about fair use, which the jury answered affirmatively, could have been supported by multiple reasons. *See id.* at 1195.

[259] *See id.* at 1206.

[260] *See id.*

Java API declaring code, was not a market substitute for Java—"Google's Android platform was part of a distinct (and more advanced) market than Java software."[261]

In addition to this more traditional review of facts relating to market substitution, the *Google v. Oracle* Court also stated that the "effect on the market" factor can encompass not only the parties' financial gains and losses but also "the public benefits the copying will likely produce."[262] This inquiry includes how the copying relates to "copyright's concern for the creative production of new expression" and the degree of "importance" of those benefits compared to the parties' monetary gains or losses.[263] The Court found that Google's copying benefitted the public because application programmers were already deeply familiar with the Java APIs.[264] Requiring programmers to learn a new set of declaring functions would allow Oracle to stifle innovation in new markets.[265]

The Court's discussion in *Google v. Oracle* of quasi-antitrust lock-in effects reflects a deeper concern raised in many of the amicus briefs about open-source norms for APIs.[266] The proprietor of an operating system, software package, or device sometimes releases APIs publicly for free.[267] This often happens when the underlying system, package, or device provides a kind of infrastructure for other applications.[268] Microsoft, for example, publicly releases APIs for its Windows operating system.[269] An operating system is subject to network effects, meaning it is more valuable to each user as more users adopt the platform.[270] Microsoft encourages the development of third-party applications that work with Windows because successful applications grow the user base and make the platform even more valuable

[261] *Id.* at 1207.

[262] *Id.* at 1206.

[263] *Id.*

[264] *See id.* at 1208.

[265] *Id.*

[266] *See, e.g.*, Brief *Amicus Curiae* of the Computer & Communications Industry Ass'n in Support of Petitioner at 4, Google v. Oracle, 141 S. Ct. 1183 (No. 18-956), https://perma.cc/4TWX-PB6Z; *see Amicus Curiae* Brief of Developers Alliance in Support of Petitioner at 11, Google v. Oracle, 141 S. Ct. 1183 (No. 18-956), https://perma.cc/S9JE-56FW.

[267] *See What Is Open Source?*, OPENSOURCE.COM, https://opensource.com/resources/what-open-source (last visited Feb. 16, 2024).

[268] *See Build Desktop Apps for Windows*, MICROSOFT: LEARN, https://learn.microsoft.com/en-us/windows/apps/desktop/#choose-your-app-type (last visited Feb. 16, 2024).

[269] *See id.*

[270] *See* Caroline Banto, *Network Effect: What It Is, How It Works, Pros and Cons*, INVESTOPEDIA (Aug. 2, 2023), https://www.investopedia.com/terms/n/network-effect.asp.

to all users.[271] The same is true of APIs for Apple operating systems.[272] Although Windows and Apple dominate the market for desktop and laptop operating systems, their open API programs facilitate flourishing application markets.

But not all APIs are open. Sometimes a proprietor keeps all APIs in-house.[273] This might be the case, for example, with a complex device that is more of a commodity than a platform, such as Tesla electric vehicles.[274] Alternatively, a proprietor might license APIs to certain business partners, as was the case with the Java APIs at issue in *Google v. Oracle*.[275] As Justice Thomas noted in his dissent in that case, other platform companies, including Amazon, had licensed Java APIs, so there was an existing market for such licenses.[276] Justice Thomas also noted that, after the litigation commenced, Google had released six versions of the Android operating system without using the Java APIs at issue, accounting for more than ninety percent of Android devices.[277] Moreover, Google itself had used its dominance in Internet searches to enhance Android's position in the mobile operating system market.[278] In contrast, both Goldsmith and Warhol were in the business of selling images to magazines and other publications as illustrations. At least according to the *Warhol* majority, Warhol's print was a market substitute for Goldsmith's photograph.[279]

[271] *See* Brief of Microsoft Corp. as *Amicus Curiae* in Support of Petitioner at 12-13, Google v. Oracle, 141 S. Ct. 1183 (2021) (No. 18-956).

[272] *See Apple Developer Documentation*, APPLE DEV., https://developer.apple.com/documentation/ (last visited Feb. 16, 2024).

[273] *Open vs. Closed APIs*, 3PILLAR GLOBAL (June 17, 2021), https://www.3pillarglobal.com/insights/open-vs-closed-apis/.

[274] Tesla has not released a public API but a group of coders is trying to publish a reverse-engineered version of a TESLA API. *See* TESLA API, https://www.teslaapi.io/ (last visited Feb. 16, 2024); Jamie Bailey, *How to Build a Tesla Data Dashboard with the Tesla API*, MEDIUM (Apr. 15, 2020), https://medium.com/initial-state/how-to-build-a-tesla-data-dashboard-with-the-tesla-api-4ebee4b9827c.

[275] *See* Google v. Oracle, 141 S. Ct. at 1210 (Thomas, J., dissenting).

[276] *Id.* at 1216.

[277] *See id.* at 1217.

[278] *Id.* at 1217-18 (citing Case AT.40099, Google Android, Comm'n Decision (July 18, 2018), https://ec.europa.eu/competition/antitrust/cases/dec_docs/40099/40099_9993_3.pdf).

[279] *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 535-36 (2023). The *Warhol* dissent persuasively argued that Warhol's aesthetic and artistic purpose differed from Goldstein's to such a degree that Warhol prints really were not market substitutes for Goldstein photos. *Id.* at 573-74 (Kagan, J., dissenting).

2. Markets for AI Training Data and Transaction Costs

So is the vast corpus of Internet content more like declaration code in an API or is it more like an image meant for publication in a magazine? The vastness of the corpus precludes any single answer. Getty Images, for example, alleged that OpenAI used material scraped from its catalog (watermarks and all) to train Dall-E.[280] Getty offers its images for a fee, and it does so for a variety of purposes.[281] Markets for AI training data are only beginning to develop, but such markets could represent a natural extension for an enterprise such as Getty that possesses the rights to millions of images that are already identified and tagged.[282] It is easy to see how OpenAI's fair use arguments might fail against Getty's claims.

Consider instead a Facebook user group for amateur astrophotographers, such as the author of this paper, who use a certain kind of telescope.[283] Users who post photos to this group do not expect any remuneration beyond some admiring "likes." Indeed, the hobby is so expensive, time-consuming, and frustrating that no one does it except for the personal satisfaction of occasionally producing an interesting picture. There is no present market for these images, either in magazines or as AI training data.[284] The same can be said for most of the photos, videos, Tik-Toks, blogs, and so on that make up the Internet's content layer. A fair use defense, therefore, seems much more robust for this content.

As discussed, however, markets for AI training data are rapidly evolving.[285] Markets for the use of ordinary web content in training data are

[280] Matt O'Brien, *Photo Giant Getty Took a Leading AI Image-Maker to Court. Now It's Also Embracing the Technology*, AP NEWS (Sept. 25, 2023, 8:46 AM), https://apnews.com/article/getty-images-artificial-intelligence-ai-image-generator-stable-diffusion-a98eeaaeb2bf13c5e8874ceb6a8ce196.

[281] *Help Center: Using Files*, GETTY IMAGES, https://www.gettyimages.com/faq/working-files (last visited Feb. 16, 2024).

[282] *See Premium Access*, GETTY IMAGES, https://www.gettyimages.com/enterprise/premium-access (last visited Feb. 16, 2024).

[283] *See Celestron RASA Owners and Imaging*, FACEBOOK, https://www.facebook.com/groups/2341262949302876/ (last visited Feb. 16, 2024).

[284] As an amateur astrophotographer, having one's image appear in a publication such as *Astronomy* magazine is a badge of honor, but the magazine does not pay for unsolicited submissions. *See Photo Submission Guidelines*, ASTRONOMY, https://www.astronomy.com/photo-submission-guidelines/ (last visited Feb. 16, 2024).

[285] *See* Sobel, *supra* note 44, at 75 ("Does training data for machine learning constitute a market that is traditional, reasonable, or likely to develop? Surprisingly, it often does."); *see also supra* Section II.C.

conceivable and likely, absent a blanket fair use rule.[286] The problem for such markets is not supply or demand—it is transaction costs. It would, of course, be impossible for an AI developer to identify and clear billions of rights claims on an individual basis. Yet this problem is not unique to AI training data. Many tried-and-true solutions have arisen to deal with the transaction costs of clearing multiple individual IP claims for traditional purposes of reproduction, distribution, and derivative works, including blanket licenses, market clearinghouses, technological measures, and compulsory licenses.[287]

3. Mitigating Transaction Costs: Market Clearinghouses and Collective Rights Management for AI Training Data

One set of possible solutions for dealing with such transaction costs involves private ordering. As noted, consent—that is, licensing—lies at the heart of how copyrighted materials are made available on the Internet.[288] Rights management organizations and market clearinghouses can aggregate rights, offer users standard license terms, and distribute revenues to rights holders based on a formula or for a set fee.

For example, performance rights societies, including ASCAP, BMI, and SESAC, allow venues to obtain performance rights licenses to large catalogs of music.[289] Similarly, patent pool organizations bundle patent rights for core technologies such as wireless networking.[290] These solutions involve well-known trade-offs: there are still organizational and administrative transaction costs built into the license fees, the organizations can become forums for horizontal price agreements and other anticompetitive behavior, and the

[286] *See* Sobel, *supra* note 44, at 75; *cf.* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 592 (1994) ("The market for potential derivative uses includes only those that creators of original works would in general develop or license others to develop."); Am. Geophysical Union v. Texaco Inc., 60 F.3d 913, 930 (2d Cir. 1994) (asking whether licensing revenue is "traditional, reasonable, or likely to be developed").

[287] DANA A. SCHERER, CONG. RSCH. SERV., R43984, MONEY FOR SOMETHING: MUSIC LICENSING IN THE 21ST CENTURY 9, 14 (2018).

[288] *See supra* Section III.A.

[289] *See BMI Services*, BMI, https://www.bmi.com/services (last visited Feb. 16, 2024); *About Music Licensing*, SESAC, https://www.sesac.com/about-music-licensing/ (last visited Feb. 16, 2024); *ASCAP Payment System*, ASCAP, https://www.ascap.com/help/royalties-and-payment/payment (last visited Feb. 16, 2024).

[290] *See, e.g.*, *Wifi 6*, SISVEL, https://www.sisvel.com/licensing-programmes/Wi-Fi/wifi-6/ (last visited Feb. 16, 2024). Patents, of course, provide different exclusive rights than copyrights—for patents, "makes, uses, offers to sell, or sells," rather than reproduction, distribution, adaptation, and the other exclusive rights of copyright. *Compare* 35 U.S.C. § 271 *with* 17 U.S.C. § 106. The concept of a rights clearinghouse, however, is similar.

bundled content or patents might include many items of dubious value.[291] Because of competition concerns, music performance rights societies are governed by antitrust consent decrees dating back to the 1940s, and patent pools usually must license on fair, reasonable, and non-discriminatory (FRAND) terms to avoid antitrust violations.[292]

Entities that serve as market clearinghouses sell content licenses through catalogs of material available *a la carte* or through bulk pricing plans. For example, Getty Images serves as a market clearinghouse for independent graphic artists, photographers, videographers, and animation designers.[293] As another example, take Netflix, Amazon Prime Video, Hulu, and other streaming services, which aggregate film and television content and deliver it to subscribers for a monthly fee.[294] Some of these services use their distribution platforms to offer sublicensing subscriptions to other content aggregators.[295] And as yet another example, take social media sites such as YouTube and TikTok that aggregate content submitted by users, including both large commercial players and individual creators.[296]

These content aggregators are less likely to face antitrust scrutiny than collective rights organizations or patent pools because the individual contributors are independent contractors or licensors who have no role in organizational governance or price setting for the organization's customers.[297] Of course, large commercial content aggregators, such as Getty

[291] *See* Robert P. Merges & Michael Mattioli, *Measuring the Costs and Benefits of Patent Pools*, 78 OHIO ST. L.J. 281, 299 (2017).

[292] Jem Aswad, *Justice Department Leaves Music Industry Consent Decrees Unchanged*, VARIETY (Jan. 15, 2021, 12:23 PM), https://variety.com/2021/music/news/justice-department-music-consent-decrees-unchanged-1234886620/; Herbert Hovenkamp, *FRAND and Antitrust*, 105 CORNELL L. REV. 1683, 1683 (2020).

[293] *See Put Your Creativity to Work*, ISTOCK BY GETTY IMAGES, https://www.istockphoto.com/workwithus (last visited Feb. 16, 2024).

[294] *See About*, NETFLIX, https://about.netflix.com/en (last visited Feb. 16, 2024); *About Hulu*, HULU PRESS, https://press.hulu.com/corporate/ (last visited Feb. 16, 2024); *Prime Video: Home*, AMAZON PRIME, https://www.amazon.com/gp/video/storefront?contentId=IncludedwithPrime&contentType=merch&merchId=IncludedwithPrime (last visited Feb. 16, 2024).

[295] *See Prime Video: Store*, AMAZON PRIME, https://www.amazon.com/gp/video/storefront/ref=atv_hm_hom_c_9zZ8D2_str?contentType=home&contentId=store (last visited Feb. 16, 2024).

[296] *See Everything You Need to Create on YouTube*, YOUTUBE CREATORS, https://www.youtube.com/creators/ (last visited Feb. 16, 2024); *Our Mission*, TIKTOK, https://www.tiktok.com/about?lang=en (last visited Feb. 16, 2024).

[297] *See TikTok Creator Fund Terms*, TIKTOK, https://www.tiktok.com/legal/page/global/tiktok-creator-fund-terms/en (last visited Feb. 16, 2024).

and Amazon, may face criticism since they are subject to network effects and can squeeze both content contributors and consumers. Getty, for example, has been criticized for selling licenses that include public domain content.[298]

But Getty also faces healthy market competition from large players, such as Adobe and Shutterstock, as well as from small competitors.[299] Presently, the stock image market is worth nearly $4 billion and is expected to grow to $7 billion over the next five years.[300] The global video streaming market, in which Netflix and Amazon Prime compete, is worth over $455 billion and is expected to grow to over $1.9 trillion by 2030.[301] Such large markets produce positive spillovers in the form of jobs, technological developments, and growth in equities markets that must be balanced against concerns about network effects and market concentration.

It is not difficult to imagine collective rights management organizations for AI training data, as that data could involve commercially available books, music, sound recordings, television programs, and films. These organizations could easily coordinate with existing media distributors, such as Amazon, the major record labels, and established film and television streaming providers. And that arrangement could extend existing business models into licenses for AI training data. Again, network effects and market concentration are major concerns—very few observers would likely be sanguine about Amazon dominating AI training. But the alternative seems to be that equally large players such as Google and Microsoft dominating AI applications with the benefit of free training material.

The examples above involve monetary licenses for commercially produced content, but the Internet's open-source ethos has always bristled at the commercialization of cyberspace. Open-source licenses, including Creative Commons licenses and the GNU Public License, have provided a mechanism through which authors could make content available for reuse

[298] *See* Mike Masnick, *Getty Images Sued Yet Again for Trying to License Public Domain Images*, TECHDIRT (Apr. 21, 2019, 9:42 AM), https://www.techdirt.com/2019/04/01/getty-images-sued-yet-again-trying-to-license-public-domain-images/.

[299] *See Stock Images And Videos Market - Global Outlook & Forecast 2023-2028*, ARIZTON (June 2023), https://www.arizton.com/market-reports/stock-images-and-stock-videos-market (webpage summarizing the report).

[300] *Id.*

[301] *Video Streaming Market Size, Share & COVID-19 Impact Analysis, [. . .] and Regional Forecast, 2023-2030*, FORTUNE BUS. INSIGHTS, https://web.archive.org/web/20230628152657/https://www.fortunebusinessinsights.com/video-streaming-market-103057 (last visited Mar. 20, 2024) (summary of report).

under noncommercial terms.[302] Such licenses can further foster the commons through "viral" terms that require adaptations to entail similar licensing terms.[303] In fact, Creative Commons is presently engaged in a process relating to AI training and applications "to consider not only the copyright system in which [Creative Commons] licenses operate, but also issues of accountability, responsibility, sustainability, cultural rights, human rights, personality rights, privacy rights, data protection, and ethics."[304]

It also is not difficult to imagine how a collective rights management organization would work on a prospective basis for the bulk of information available on the Internet, much of which is contributed by individuals to social media sites.[305] Individuals who contribute content through social media sites, such as YouTube, TikTok, Instagram, Facebook, and LinkedIn, typically retain the copyrights to that content and agree to terms of service regarding how the content can be used.[306] These terms could include provisions about whether the content could be made available as AI training data. The platforms could work out some kind of revenue sharing model with users, depending on how markets develop. And as organizations such as Creative Commons develop noncommercial license terms, it will become easy for individuals and organizations to make their materials available as training data for free in a commons-forward viral licensing model.

In other words, concerning most commercially available content and individually contributed noncommercial Internet content, infrastructure already exists for markets in AI training data.

[302] *See What We Do*, CREATIVE COMMONS, https://creativecommons.org/about/ (last visited Feb. 16, 2024); *GNU General Public License*, GNU OPERATING SYS. (June 29, 2007), https://www.gnu.org/licenses/gpl-3.0.en.html.

[303] *See What We Do*, *supra* note 302.

[304] Brigitte Vézina & Sarah Hinchliff Pearson, *Should CC-Licensed Content Be Used to Train AI? It Depends.*, CREATIVE COMMONS (Mar. 4, 2021), https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/; *see also 2023 CC Global Summit: Registration, Call for Proposals, and Scholarships Now Open*, CREATIVE COMMONS (June 2, 2023), https://creativecommons.org/2023/06/02/2023-cc-global-summit-registration-call-for-proposals-and-scholarships-now-open/.

[305] *See* Stacy Jo Dixon, *Number of Social Media Users Worldwide from 2017 to 2027*, STATISTA (Aug. 29, 2023), https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ ("In 2022, over 4.59 billion people were using social media worldwide . . . .").

[306] *See* McCabe Curwood, *Who Owns My Social Media Content?*, LEXOLOGY (May 16, 2017), https://www.lexology.com/library/detail.aspx?g=a2627dc8-1d2a-4a2a-ae48-04f3f0cc2815.

4. Compulsory Licenses for AI Training Data

Although private ordering seems quite feasible, there is potentially a set of legal and market barriers to private ordering solutions for aggregating copyrighted material as AI training data. First, under U.S. law, a non-exclusive licensee does not have standing to sue for copyright infringement.[307] Therefore, an aggregator or social media site might not be able to protect a market for collected, copyrighted training data. The 1976 Act, however, allows the divisibility of the bundle of exclusive rights.[308] If a licensor conveys to an agent the exclusive right to grant sublicenses, the grant is considered "exclusive" under the 1976 Act, even if the licensor retains the right to grant other licenses.[309] This is how stock photography providers, for example, can enforce rights against third parties.[310] But such agreements must be carefully structured to ensure that the licensee actually receives at least some exclusive grant of copyright—such as an exclusive grant to make sublicenses for certain purposes—or else the licensee will not have standing to sue third parties.[311] This would require many aggregators and social media sites to obtain stronger copyright licenses than they presently obtain from

[307] *See* 17 U.S.C. § 501(b) (allowing only "[t]he legal or beneficial owner of an exclusive right under a copyright" to bring an infringement action); 3 MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 12.02[B][1] (2024); Minden Pictures, Inc. v. John Wiley & Sons, Inc., 795 F.3d 997, 1003 (9th Cir. 2015).

[308] 17 U.S.C. § 201(d).

[309] *Minden Pictures, Inc.*, 795 F.3d at 1004. Some case law suggests that agreements must be carefully structured to ensure that the licensee in fact receives at least some exclusive grant of copyright—such as an exclusive grant to make sub-licenses for certain purposes—or else the agent / licensee will not have standing to sue third parties. *See* Creative Photographers, Inc. v. Julie Torres Art, LLC, No. 1:22-CV-00655, 2023 WL 2482962, at *5 (N.D. Ga. Mar. 13, 2023) (holding exclusive art agency agreement that did not clearly convey copyright interest insufficient for standing to sue for copyright infringement); Greg Young Publ'g, Inc. v. Zazzle, Inc., No. 2:16-CV-04587, 2017 WL 2729584, at *2 (C.D. Cal. May 1, 2017) (holding "exclusive representative" for negotiating licenses had no standing to sue for copyright infringement where representation agreement did not convey rights under copyrights).

[310] *See Minden Pictures, Inc.*, 795 F.3d at 1004.

[311] *See Creative Photographers, Inc.*, 2023 WL 2482962, at *5; *Greg Young Publ'g, Inc.*, 2017 WL 2729584, at *2.

users.[312] Other aggregators, such as Getty Images, already offer tiers in which contributors can be non-exclusive or exclusive contributors.[313]

Second, some of the major players might be uninterested in facilitating user rights for anticompetitive reasons. For example, YouTube is owned by Google, which has acquired at least thirty AI startup companies worth almost $4 billion since 2009.[314] Google might wish to consume AI training data from user content on YouTube and its other sites for free, either under a claim of fair use or as a condition of its terms of service, while selectively asserting contract or tort-based claims against those who mine data from its sites. The large entities with huge troves of user content and vested interests in AI—including the GAMAM companies—may not want a competitive market for this kind of use, even if they could profit from brokering the data to third parties.[315] Their interest in controlling AI development might outweigh whatever profits they could make from brokering training data to other developers.

In response to such concerns and as a backstop to private ordering solutions, the 1976 Act could encode a compulsory license for AI training data. There are already numerous compulsory licenses in the law, including for sound recordings of musical works, noncommercial broadcasting, satellite retransmission, cable system retransmission, and digital audio transmission.[316] As the subject matter suggests, compulsory licensing is a common solution to copyright challenges presented by disruptive technologies that give rise to new industries.[317] Compulsory licenses can be

[312] *See, e.g.*, *Terms of Service: Your Content and Conduct*, *supra* note 130 (stating that YouTube users grant "a worldwide, non-exclusive, royalty-free, sublicensable and transferable license to use that Content (including to reproduce, distribute, prepare derivative works, display and perform it) in connection with the Service").

[313] *See Put Your Creativity to Work*, GETTY IMAGES, https://www.gettyimages.com/workwithus (last visited Feb. 16, 2024).

[314] Aaron Hurst, *Google Revealed to Have Acquired the Most AI Startups Since 2009*, INFO. AGE (Feb. 18, 2020), https://www.information-age.com/google-revealed-acquired-most-ai-startups-since-2009-15415/.

[315] The GAMAM companies are Google, Apple, Microsoft, Amazon, and Meta. J. Clement, *Google, Amazon, Meta, Apple, and Microsoft (GAMAM) - Statistics & Facts*, STATISTA (Jan. 10, 2024), https://www.statista.com/topics/4213/google-apple-facebook-amazon-and-microsoft-gafam/#topicOverview.

[316] 17 U.S.C. §§ 111, 115, 118-119, 122.

[317] For example, early cable television systems began as a hacker's project in the late 1940s, using hilltop antennas connected to coaxial cable to distribute broadcast television signals in areas with bad reception. *See* Matthew G. Anderson, *Wired: Cable TV's Unlikely Beginning*, PA. CTR. FOR THE BOOK, https://pabook.libraries.psu.edu/literary-cultural-heritage-

difficult to administer and are subject to criticisms that the terms quickly become outdated and unfair.[318] Royalty calculations can also become overly complex. For example, licensors may overpay or licensees may be underpaid, technology may outpace the system, and the entrenched system may stifle the growth of new technologies and markets.[319] But a compulsory license could serve as a background rule and norm that encourages creative private ordering solutions.

5. Technological Measures for Rights Management

A final set of transaction cost and enforcement cost problems with large-scale collective rights management involves technological measures. Collective rights clearance for AI training data might require a protocol that is more robust than Google's robots.txt file, both as a technological barrier to unauthorized crawling and scraping, and as a permissions and accounting mechanism for authorized crawling and scraping.[320] The protocol must define

map-pa/feature-articles/wired-cable-tvs-unlikely-beginning (last visited Mar. 20, 2024); Brad Adgate, *The Rise and Fall of Cable Television*, FORBES (Nov. 2, 2020, 4:09 PM), https://www.forbes.com/sites/bradadgate/2020/11/02/the-rise-and-fall-of-cable-television/?sh=1dfaea5e6b31.

As the practice of connecting antenna to cable began to grow into an industry, FCC regulation and copyright challenges spurred by television and movie studios mounted. REG. OF COPYRIGHTS, *supra* note 113, at i-ii. In 1968, the Supreme Court held that extending local broadcast signals from antennas through cable wires was not a "performance" of a work under the then-extant Copyright Act of 1909. Fortnightly Corp. v. United Artists Television, Inc., 392 U.S. 390, 402 (1968). In 1974, the Court extended this holding to the reception of "distant" broadcast signals. Teleprompter Corp. v. Columbia Broad. Sys., Inc., 415 U.S. 394, 414 (1974). Meanwhile, the FCC began to issue regulations attempting to facilitate the growth of this new technology and industry while recognizing the interests of content creators—the broadcasting companies—that the Court had held were not anticipated in the 1909 Act. REG. OF COPYRIGHTS, *supra* note 113, at ii-iii. The FCC rules allowed some broadcasters to obtain exclusivity over some programming that the cable operators were not allowed to carry. *Id.* at iii. Finally, in the 1976 Act, Congress reached a compromise that made cable-television retransmission an infringement but that established a compulsory-licensing regime. *See id.* at iv; 17 U.S.C. § 111(c).

[318] *See, e.g.*, Dylan Smith, *Is It Time to Repeal the Section 115 Compulsory License? One Songwriter Is Formally Urging the Copyright Office to Do Just That*, DIGIT. MUSIC NEWS (June 23, 2023), https://www.digitalmusicnews.com/2023/06/23/section-115-compulsory-license-repeal-george-johnson/.

[319] *See id.*; REG. OF COPYRIGHTS, *supra* note 113, at ix-xiii.

[320] *See* Transcript of Proceedings, *supra* note 2, at 31:12-18 (noting that technology similar to robots.txts "is there and some of the concerns can be abated if these tools become mandated or just widely used"); *id.* at 39:8-10 ("We see metadata and CMI as key to being

permitted crawling and scraping, such as for search indexing, and be difficult to circumvent.

The robots.txt protocol was, in fact, updated in 2022—the first update since its creation in 2004.[321] As Google's own instructions make clear, however, the robots.txt file is far from foolproof.[322] Indeed, commercial web crawling and scraping service providers openly brag about how they avoid web scraping blocks and bans from the robot.txt file and other sources.[323]

A more robust robots.txt-like protocol could be supported by provisions in the 1976 Act concerning "copyright management information" ("CMI"). The 1976 Act presently makes it unlawful to intentionally remove or alter CMI.[324] An injured party can recover actual or statutory damages against a party who distributes copies of works while knowing that CMI "has been removed or altered without authority of the copyright owner" if the defendant had "reasonable grounds to know, that it will induce, enable, facilitate, or conceal an infringement" of copyright.[325] As many commentators have suggested, policymakers could extend the definition of CMI, either by statutory amendment or by regulation through the Register of Copyrights, to include permissions regarding use for AI training data.[326]

able to protect artists' works in an AI environment . . . ."); *id.* at 40:11-41:9 (discussing open content management standards being developed by Adobe and others "that will give artists the ability and tools . . . to identify whether they want to participate or don't want to participate and encouraging that kind of proactivity among the companies that are developing this technology to give artists a tool to control their creative work"); *id.* at 43:14-44:7 (stating that Stability AI supports protocols like robots.txt for consent to automated data aggregation).

[321] *See* Martijn Koster et al., *Robots Exclusion Protocol*, DATATRACKER (Sept. 12, 2022), https://datatracker.ietf.org/doc/rfc9309/.

[322] *See Introduction to Robots.txt*, GOOGLE SEARCH CENT. https://developers.google.com/search/docs/crawling-indexing/robots/intro (last visited Feb. 16, 2024) (noting that a robots.txt file will not necessarily prevent a page from showing up in search results).

[323] *See* Colm Kenny, *How to Avoid Web Scraping Blocks and Bans*, ZYTE (May 18, 2022), https://www.zyte.com/blog/scraping-blocks-and-bans/; Akshay Philar, *How to Manage Bans and Get Data with Zyte Data API Smart Browser*, ZYTE (Sept. 7, 2021), https://www.zyte.com/blog/manage-bans-and-get-your-data-zyte-data-api/.

[324] 17 U.S.C. § 1202(b)(1).

[325] *Id.* § 1202(b).

[326] *See id.* § 1202(c)(8) (stating that the Register of Copyrights can specify information included under the definition of CMI). In her comments at the May 2023 Copyright Office listening session, Rebecca Blake of the Graphic Artists Guild suggested that the scienter requirement in section 1202 must be modified to facilitate metadata and CMI in a training-data-permissions protocol. Transcript of Proceedings, *supra* note 2, at 38:23-39:11. This is probably correct if section 1202 is otherwise left as-is. If the definition of CMI is modified to include permission protocols for training data, such an amendment might not be necessary,

6. Markets and Technological Exceptionalism: Where Does AI Fit in the Story?

At the dawn of the Internet era in the early 1990s, some commentators argued that copyright and other traditional legal domains should be radically altered.[327] The Internet was something new, something that should be left as free as possible to grow organically from the ground up. But this kind of Internet exceptionalism was challenged from the beginning.

Internet law responded to these tensions in various ways. Section 230 of the Communications Decency Act, part of the Telecommunications Act of 1996, exempted Internet hosting sites from publisher liability.[328] This exemption garnered the praise of many open Internet activists (even though John Perry Barlow still disapproved of the CDA),[329] and it supported countless instances of learning and creativity, but it also helped produce today's toxic social media culture.[330]

The Digital Millennium Copyright Act of 1998 included a provision that riled open Internet and open-source advocates.[331] It also included safe harbors from secondary copyright liability for sites that took certain steps to limit infringing content.[332] Open Internet activists welcomed the safe harbors but raised concerns that they were insufficiently attentive to fair use.[333] Like section 230, the DMCA safe harbors have supported the dynamic creativity of Web 2.0 but have facilitated platform consolidation and cultures of abuse.[334]

because the protocol, which would be machine-readable, would itself provide actual notice. It could be helpful, though, to make clear by statutory amendment or regulation that willful blindness through failure to access the protocol is no defense.

[327] *See, e.g.*, David R. Johnson & David Post, *Law and Borders—the Rise of Law in Cyberspace*, 48 STAN. L. REV. 1367 (1996).

[328] 47 U.S.C. § 230(c).

[329] *Section 230*, ELEC. FRONTIER FOUND., https://www.eff.org/issues/cda230 (last visited Feb. 16, 2024); Richard Bennett, *The Legacy of Barlow's Cyberspace Declaration of Independence*, AM. ENTER. INST. (Feb. 10, 2016), https://www.aei.org/technology-and-innovation/telecommunications/legacy-barlows-cyberspace-declaration-independence/.

[330] *See* Michael D. Smith & Marshall Van Alstyne, *It's Time to Update Section 230*, HARVARD BUS. REV. (Aug. 12, 2021), https://hbr.org/2021/08/its-time-to-update-section-230.

[331] *Digital Millennium Copyright Act*, ELEC. FRONTIER FOUND., https://www.eff.org/issues/dmca (last visited Feb. 16, 2024).

[332] 17 U.S.C. § 512.

[333] *Digital Millennium Copyright Act*, *supra* note 331.

[334] "Web 2.0" is term coined in the late 1990s for a World Wide Web that emphasizes user-generated content. *See* Kinsa Yazar, *Web 2.0*, TECHTARGET, https://www.techtarget.

Acting under the deregulatory impetus of section 706 of the Telecommunications Act, the FCC's decision in 2002 to classify broadband Internet as an "information service" ensured a light regulatory touch rather than the heavy one imposed on other telecommunications services.[335] Although the Internet never became the libertarian utopia imagined by Barlow, this light touch regulation allowed technologists and community members to manage broadband Internet's growth rather than bureaucrats.

But progressives changed their minds when the Internet backbone market became highly concentrated.[336] Leaving cyberspace to the people turned out to mean leaving cyberspace's physical backbone to a few large corporations. It became obvious that cyberspace was not a borderless world after all.[337] As a result, progressives advocated for network neutrality rules in the FCC's

com/whatis/definition/Web-20-or-Web-2?Offer=abt_pubpro_AI-Insider (last updated Jan. 2023). YouTube is a good example of this double effect. The platform has grown exponentially and offers a vast array of informational and entertainment content. But YouTube has been criticized by users for enforcing the DMCA notice-and-takedown rules too aggressively in favor of large commercial interests and by artists for allowing widespread "piracy" to occur on the site. *See, e.g.*, Sarah Clough-Segall, *YouTube's Copyright Policy: Pitfalls Aplenty for Video Creators*, JDSUPRA (Oct. 13, 2020), https://www.jdsupra.com/legalnews/youtube-s-copyright-policy-pitfalls-23119/ ("YouTube's current copyright procedures are laden with pitfalls which deter content creators from creating and posting new work"); *cf.* Maria Schneider, *What Do Whore Houses, Meth Labs, and YouTube Have in Common?*, MUSIC TECH. POL'Y BLOG (Sept. 27, 2016), https://musictechpolicy.com/2016/09/27/guest-post-by-schneidermariawhat-do-whore-houses-meth-labs-and-youtube-have-in-common/. Schneider attempted to lead a class action against YouTube for alleged failure to take down infringing works but the case was voluntarily dismissed a day before trial after the court refused to certify a class. *See* Stuart Dredge, *Maria Schneider's YouTube Lawsuit Dismissed Just Before Trial*, MUSIC ALLY (June 13, 2023), https://musically.com/2023/06/13/maria-schneiders-youtube-lawsuit-dismissed-just-before-trial/.

[335] Inquiry Concerning High-Speed Access to the Internet Over Cable & Other Facilities, 17 FCC Rcd. 4798, 4802 (2002).

[336] Protecting and Promoting the Open Internet, 30 FCC Rcd. 5601, 5920 (2015).

[337] *See* JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD 10 (2006). Wu coined the term "network neutrality" and was one of the key advocates of network-neutrality rules. *See, e.g.*, Tim Wu, A Proposal for Network Neutrality (June 2002), http://www.timwu.org/OriginalNNProposal.pdf; Tim Wu, *Network Neutrality, Broadband Discrimination*, 2 J. ON TELECOMM. & HIGH TECH. L. 141 (2003); Chaim Gartenberg, *Tim Wu, the 'Father of Net Neutrality,' Is Joining the Biden Administration*, THE VERGE (Mar. 5, 2021, 9:49 AM), https://www.theverge.com/2021/3/5/22315224/tim-wu-net-neutrality-antitrust-big-tech-biden-administration-national-economic-council.

2015 Open Internet Order.[338] These rules, however, were quickly reversed after the FCC's composition changed during the Trump administration.[339]

As these examples suggest, the Internet is both exceptional and ordinary. Today, enormous problems relating to cybercrime, surveillance, intellectual property, equal access, harassment, and privacy continue to bedevil cyberspace.[340] The same mix of exceptional and ordinary will characterize AI law and policy but at even greater speed and scale.

This rapid change implicates copyright policy. In the discussion of the transformativeness fair use factor in *Warhol*, the majority instructed courts to ask whether the new use "furthers the goal of copyright, namely, to promote the progress of science and the arts, without diminishing the incentive to create."[341] The instinct of some scholars, technologists, and policymakers immersed in the culture of Internet exceptionalism is to remove copyright as a potential speedbump through fair use.[342] But while some AIs may serve to promote science and the arts, others may not. Indeed, it is possible that AIs could severely degrade or destroy science, the arts, and other human endeavors.[343] The uncertainty combined with the scale of change counsels against any generalized fair use rules like non-expressive use.[344]

Because humanity stands on the threshold of the next technological revolution, no one argues for AI exceptionalism against regulation. Perhaps some lessons were learned from the excesses of early Internet

[338] *See generally* Protecting and Promoting the Open Internet, 30 FCC Rcd. 5601.

[339] In its 2015 "Open Internet Order" imposing network neutrality rules, the FCC decided to reclassify broadband Internet service as telecommunications services under Title II, with regulatory forbearance of certain other rules. *Id.* at 5686. The Open Internet Order subsequently was reversed in the 2018 "Restoring Internet Freedom Order." Restoring Internet Freedom, 33 FCC Rcd. 311, 509 (2018).

[340] *See, e.g.*, David W. Opderbeck, *Cybersecurity and Data Breach Harms: Theory and Reality*, 82 MD. L. REV. 1001 (2023).

[341] Andy Warhol Found. For the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 531 (2023).

[342] *See* Sobel, *supra* note 137.

[343] As Professor Gary Marcus testified at a May 2023 congressional hearing on AI safety, "We have built machines that are like bulls in a china shop—powerful, reckless, and difficult to control." *Oversight of A.I.: Rules for Artificial Intelligence: Hearing Before the S. Judiciary Comm. Subcomm. on Privacy, Tech. and the Law*, 118th Cong. (2023) (testimony of Gary Marcus).

[344] This is a restatement of the "precautionary principle" often used in environmental and public health ethics. *See* David Kriebel et al., *The Precautionary Principle in Environmental Science*, 109 ENV'T HEALTH PERSPS. 871 (2001).

exceptionalism. In fact, AI industry leaders, are *asking* for regulation.[345] We might question the sincerity and motives of these requests, but certainly there is no one declaring the independence of AI—except, perhaps, its independence from copyright.[346]

While copyright should not drive AI policy, a copyright speedbump might create spillover benefits for AI policy. One of these benefits is privacy. Consider a parent who uploads video clips of a child's birthday party on Instagram or TikTok.[347] Ordinarily, there is not a market for these videos. The poster may intend to share the videos with friends and family or to get some "likes" from the broader social community. But these videos involve serious privacy concerns. Undoubtedly, the parent chose to make clips of the child publicly available for others to view, and the parent did so regardless of whether the decision was wise or made with a full understanding of the site's privacy policies and controls. It seems unlikely, however, that the parent would have wanted the child's face to be used to train someone else's AI. Here, if new market mechanisms could restrict the use of videos such as these for AI training purposes, then such mechanisms could promote both the dynamic competition purposes of copyright and privacy values by giving the parent more control over how the clips are used.

A second spillover benefit relates to data quality. Take ChatGPT, for example. It is the uber digital native. It was born and raised on the web, and initiated into both the web's light and dark sides.[348] Sadly, though

[345] *See* Courtney Rozen, *AI Leaders Are Calling for More Regulation of the Tech. Here's What That May Mean in the U.S.*, WASH. POST (May 31, 2023, 12:26 PM), https://www.washingtonpost.com/business/2023/05/31/regulate-ai-here-s-what-that-might-mean-in-the-us/770b9208-ffd0-11ed-9eb0-6c94dcb16fcf_story.html; *Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI*, THE WHITE HOUSE (July 21, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

[346] As Sobel correctly notes, "Commercial machine learning, trained on expressive media, promises tremendous social value. But it is not the sort of value that fair use exists to foster. Unlike the benefits realized by, say, scholarship, the value of advanced machine learning is internalized by the large firms that furnish those services." Sobel, *supra* note 44, at 89.

[347] *See, e.g.*, @Kelle.cross, TIKTOK (Apr. 9, 2023), https://www.tiktok.com/@kellee.cross/video/7220054164548832558?q=birthday%20party&t=1690232329512; Yoadan Ephrem Tadesse (@yoadan_ephrem), INSTAGRAM (Dec. 28, 2021), https://www.instagram.com/p/CYCF2XtqUld/.

[348] *See* Michael Conklin, *Is AI Friend or Foe: Legal Implications of Rapid Artificial Intelligence Adoption*, 26 ATL. L.J. 2, 3 (2023).

unsurprisingly, ChatGPT can become predatory, racist, antisocial, dishonest, and casually violent.[349] Because of these concerns, in March of 2023, a group of technologists, academics, and business and policy leaders issued a letter calling for a moratorium on some AI development and research.[350] Their recommendations included restricting access to certain kinds of computing power "[t]o prevent reckless training of the highest risk models."[351] Restricting computing power is likely an unwise and unrealistic goal under U.S. law since the computer power in question is privately owned.[352] But restricting access to the Internet's dark reaches is quite feasible through copyright law.

Of course, if there are commercial licensing mechanisms for user content, there is no guarantee that the data made available under such licenses will be good data. If there is a general compulsory licensing mechanism, all the bad data will still be available as well. But the *cost* of data will limit recklessness because users will be unwilling to pay for bad data. A cost mechanism would accelerate markets for entities that specialize in cleaning and tagging data sets. Data brokerages could acquire content from trusted individuals and repackage it for training, similar to present crowd-sourced models but on a far greater scale. Licensing costs would thereby internalize the externalities of models produced with bad data while also producing positive spillovers in new data quality industries.

The effect on the market factor, then, could weigh against fair use, even for non-commercial content that trains AIs, particularly in light of the role

[349] *See, e.g.*, Kyle Wiggers, *Researchers Discover a Way to Make ChatGPT Consistently Toxic*, TECHCRUNCH (Apr. 12, 2023, 8:00 AM), https://techcrunch.com/2023/04/12/researchers-discover-a-way-to-make-chatgpt-consistently-toxic/; Prashnu Verma & Will Oremus, *ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused*, WASH. POST (Apr. 5, 2023, 2:07 PM), https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/; Sam Biddle, *The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques*, THE INTERCEPT (Dec. 8, 2022, 1:44 PM), https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/.

[350] *Pause Giant AI Experiments: An Open Letter*, FUTURE LIFE INST. (Mar. 22, 2023), https://futureoflife.org/open-letter/pause-giant-ai-experiments/. Again, it is fair to express some skepticism about the motives of some of the signatories. Do we think Elon Musk is always a rational, ethical actor? Did Getty Images sign on because of authentic concerns for the commonweal or because generative AI threatens its business model? Nevertheless, the list of signatories is extensive and their concerns are weighty.

[351] FUTURE OF LIFE INST., POLICYMAKING IN THE PAUSE 8 (2023), https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf.

[352] Prohibiting Google, Amazon, or Microsoft, for example, from using their own computing facilities could comprise a regulatory taking worth billions of taxpayer dollars. *See* U.S. CONST. amend. V; Pa. Coal Co. v. Mahon, 260 U.S. 393, 415 (1922).

copyright may play in connection with AI policy. At the same time, the question of how AI relates to *human* sciences and arts leads to a more basic question in AI ethics and policy, which also looms behind the instinct that copyright should *not* impede AI training: could or should an AI itself have rights? This question raises a copyright concern that is more fundamental than a novel theory of non-expressive use: does training data *educate* an AI? If so, is there an argument for educational fair use? The issue is addressed in Part V.

V. Copyright and the Education of Humans and Artificial Agents

A personal anecdote illustrates the concerns raised in this Part. In a conversation with other cyber law scholars, the author of this Article expressed his opinion that courts should not apply a blanket fair use exception for AI training data. A colleague responded, "I have spent my whole life processing data and repurposing it in new works. Oops. I guess I should have been paying compulsory licenses."[353] It was a humorous and somewhat tongue-in-cheek response, but it demonstrates the instinct that AI learning is analogous to human learning. This instinct underlies arguments about non-expressive use for AI training data. As Curt Levy, President of the Committee for Justice, stated in the recent U.S. Copyright Office listening session about copyright and AI training data:

> [T]he neural networks at the heart of AI, learn from very large numbers of examples, and at a deep level, it's analogous to how human creators learn from a lifetime of examples. And we don't call that infringement when a human does it, so it's hard for me to conclude that it's infringement when done by AI.[354]

[353] Emails on file with author.

[354] Transcript of Proceedings, *supra* note 2, at 21:7-13. The Committee for Justice is a conservative think tank. *See About the Committee for Justice*, COMM. FOR JUST., https://www.committeeforjustice.org/about (last visited Feb. 16, 2024). Levy's testimony illustrates how advocates on the "copyleft" and some on the political right are making strange bedfellows around the open access to copyrighted materials for AI training. Levy further noted that:

> [t]he human brain consists of neurons connected by synapse of various strength. So, when a human sees an example, those synaptic strengths are slightly modified. . . . Neural networks consist of artificial neural networks connected by artificial synapses. When the AI is shown an example, the synaptic strengths or weights are slightly modified . . . and we call that learning.

Id. at 52:21-53:6.

From the dawn of the Internet era until today, the energy around open source, open access, open data, and the information commons has been about human learning and development—and understandably so.[355] If training an AI is analogous to educating a human being, deeper copyright concerns arise, and a broader non-expressive use principle might be appropriate. If not, the instinct is mistaken.

A. Education and the Ethics of Copyright

We use words like "train," "training data," and "learning" to describe what is happening when an AI ingests information to build models. In other words, we are educating AIs. From the earliest days of Anglo-American copyright, education has been recognized as a value that limits the scope of the property right.[356] This value arose through English law's tolerance for abridgements.[357] In eighteenth-century England, it was common practice for publishers to produce abridged versions of lengthier works to make the ideas of the underlying works available to the broader public.[358] Samuel Johnson, a great literary celebrity of the era, described the purpose of abridgements in his unpublished 1739 manuscript "Considerations on the Case of Dr T.—s Sermons Abridg'd by Mr. Cave":

> The Design of an Abridgement is to benefit mankind by facilitating the attainment of knowledge, and by contracting arguments, relations, or descriptions, into a narrow compass, to convey instruction in the easiest method without fatiguing the attention burdening the memory, or impairing the health of the Student.[359]

Johnson acknowledged that abridgment might lessen the economic value of the underlying work but asserted that "the advantage received by mankind from the easier propagation of knowledge" outweighed such a burden.[360]

[355] *See, e.g.*, Carroll, *supra* note 109, at 907.

[356] *Id.* at 963.

[357] *See* Matthew Sag, *The Prehistory of Fair Use*, 76 BROOK. L. REV. 1371, 1375 (2011).

[358] *Id.* at 1384.

[359] Samuel Johnson, *Considerations on the Case of Dr T.—s Sermons Abridg'd by Mr Cave (1739)*, THE YALE DIGIT. EDITION OF THE WORKS OF SAMUEL JOHNSON 47, 54 ¶ 19, https://web.archive.org/web/20230323054437/http://www.yalejohnson.com/frontend/sda_viewer?n=112220 (last visited Mar. 20, 2024). We might view Johnson's draft arguments with a cynical eye, since he was hired by the publisher in anticipate of a lawsuit by the holder of the copyright in Rev. Trapp's sermons. *Id.* at 48.

[360] *Id.* at 54 ¶ 20.

The seemingly absolute exception for abridgement in some early English copyright cases did not directly carry over into later American copyright law.[361] In *Folsom v. Marsh*, a seminal case discussing the American fair use doctrine, the Court found that an abridgement of the complete works of George Washington infringed the original publisher's copyright.[362] In evaluating the case in equity for injunctive relief, the Court suggested several factors to determine whether a quotation or abridgment was unlawful: "the nature and objects of the selections made, the quantity and value of the materials used, and the degree in which the use may prejudice the sale, or diminish the profits, or supersede the objects, of the original work."[363]

The *Folsom* factors informed the four fair use factors in the 1976 Act, which emphasize that teaching, scholarship, and research are potential examples of fair use.[364] The "purpose and character of the use" factor suggests that "nonprofit educational purposes" would tip the scales towards fair use.[365] Copyright's emphasis on education is evident in specific statutory exemptions for libraries, archives, and online teaching.[366] Of course, the library and archival exemptions and the TEACH Act are limited, and educational uses are evaluated under the four factors like other uses. Indeed, where markets exist for libraries to license books and other educational materials, a publisher that seeks to evade those markets likely will not have a fair use defense.[367]

Education remained an important consideration in the developing concept of fair use internationally. The Berne Convention of 1886, which the United States did not initially join, included a specific exemption for free uses "to the extent justified by the purpose, of literary or artistic works by way of illustration in publications, broadcasts or sound or visual recordings for teaching, provided such utilization is compatible with fair practice."[368] The

[361] *See* Sag, *supra* note 357, at 1374.

[362] 9 F. Cas. 342, 349 (D. Mass. 1841) (No. 4901).

[363] *Id.* at 348; 17 U.S.C. § 107.

[364] 17 U.S.C. § 107.

[365] *See id.*

[366] *See* 17 U.S.C. § 107-108, 110.

[367] *See* Hachette Book Grp., Inc. v. Internet Archive, 664 F. Supp. 3d 370, 390-91 (S.D.N.Y. 2023). The Internet Archive makes full text copies of e-books available for free. The district court distinguished *HathiTrust* and *Google Books* because, in *HathiTrust*, full copies were only available to print-disabled patrons for whom there was no established market, and in *Google Books*, only snippets were publicly available. *Id.* at 381.

[368] Berne Convention for the Protection of Literary and Artistic Works, Sept. 6, 1886, art. 10(2), S. TREATY DOC. NO. 99-27, 1161 U.N.T.S. 3.

United States, at least partially, came into compliance with the Berne Convention by adopting the 1976 Act.

Education is also a value deeply embedded in international law in relation to proprietary rights. Article 26 of the Universal Declaration of Human Rights states that "[e]veryone has the right to education."[369] Moreover, article 27(1) states that "[e]veryone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits," while article 27(2) provides that "[e]veryone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author."[370] The World Summit on the Information Society ("WSIS") Declaration of Principles states, "We recognize that education, knowledge, information and communication are at the core of human progress, endeavour and well-being."[371] To this end, the WSIS Declaration of Principles argues that "[a] rich public domain is an essential element for the growth of the Information Society, creating multiple benefits such as an educated public, new jobs, innovation, business opportunities, and the advancement of sciences."[372]

In the past, while human beings knew that the brain had some role in cognition, they often assigned higher levels of understanding to the soul, spirit, or mind as a kind of nonmaterial property.[373] In recent decades, scientists have discovered that human learning is facilitated by physical and chemical connections among neurons.[374] Philosophers, scientists, and theologians now debate over whether such higher levels of cognition really can be reduced entirely to the material structure and chemistry of the brain.[375] Indeed, this is the essential question in the debate over whether AGI is possible. Human cognition may be ineluctably tied to a human body. Or

[369] G.A. Res. 217 (III) A, Universal Declaration of Human Rights, art. 26 (Dec. 10, 1948).

[370] *Id.* art. 27.

[371] World Summit on the Information Society, Declaration of Principles: Building the Information Society: A Global Challenge in the New Millennium, art. 8, Doc. No. WSIS-03/GENEVA/DOC/4-E (Dec. 12, 2003), https://www.itu.int/net/wsis/docs/geneva/official/dop.html.

[372] *Id.* art. 26.

[373] DAVID W. OPDERBECK, THE END OF THE LAW? THEOLOGY, NEUROSCIENCE, AND THE SOUL 96 (2021) [hereinafter OPDERBECK, THE END OF THE LAW?].

[374] *See* Paul S. Davies & Peter A. Alces, *Neuroscience Changes More Than You Can Think*, 2017 U. ILL. J.L. TECH. & POL'Y 141, 153 (reviewing OWEN JONES ET AL, LAW AND NEUROSCIENCE (1st ed. 2014)).

[375] For my contribution to this debate, see OPDERBECK, THE END OF THE LAW?, *supra* note 373.

human cognition may be finally irreducible and inscrutable at any precise level of detail. Maybe the human mind, like a very complex AI, is finally a black box. In any event, human learning from copyrighted materials involves biochemical reproduction, which includes transitory copies of information when it is ingested, longer term copies of things committed to memory, and the storage of chemical algorithmic tokens representing patterns and decision points.[376]

Perhaps copyright law's exclusion of these biochemical functions from the definition of reproduction is based on a faulty, prescientific philosophy of mind. In 1908, in *White-Smith Music Publishing Co. v. Apollo Co.*, the Supreme Court held that a player-piano roll was not a copy of the music inscribed on it because the music is perceived by the ear.[377] In the Copyright Act of 1909, Congress changed the *White-Smith* rule by creating the mechanical license, the forerunner of today's detailed rules about compulsory licenses for nondramatic musical works.[378] If we could describe human learning with the same level of molecular detail as we could describe the pattern of holes in a piano player roll or the lines of computer source code, human memory, too, could be considered a form of copyright reproduction. Perhaps future copyright law will entail a compulsory license merely for reading.

To state such a possibility is to recognize that it is absurd. Even the most hard-core reductive materialist in the philosophy of mind would be unlikely to equate a child's learning from a book with an unlicensed reproduction or derivative work. Our ethical intuitions and beliefs tell us that human beings are not commodities and that copyright law does not extend to how they learn from sources that are otherwise lawfully reproduced and distributed. The biochemical functions of the human brain as a limit on copyright runs deeper than merely a pragmatic or utilitarian concern. Should the same logic apply to the education of AIs? The answer depends on AI's place in society and on whether an AI could have rights analogous to human rights.

[376] *See* Joshua C. Liderman, Note, *Changing the Channel: The Copyright Fixation Debate*, 36 RUTGERS COMPUT. & TECH. L.J. 289, 289 (2010).

[377] 209 U.S. 1, 17 (1908).

[378] *See* 17 U.S.C. § 115; Copyright Act of 1909, Pub. L. No. 60-349, § 1(e), 35 Stat. 1075, 1075.

B. AI Ethics and Three Perspectives on Machine Ethics

Some notions within AI ethics suggest that AIs cannot possess anything like human rights. Instead, AIs are tools that serve humans.[379] The highly regarded Asilomar AI Principles, for example, provide that "[t]he goal of AI research should be to create not undirected intelligence, but beneficial intelligence" for humans.[380] Other representative statements in the Asilomar Principles include:

> 10) Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
>
> 11) Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity. . . . .
>
> 14) Shared Benefit: AI technologies should benefit and empower as many people as possible.
>
> 15) Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.[381]

The Asilomar Principles thus envision AI systems as tools or servants of humans. Similarly, the Ethics Guidelines for Trustworthy AI, produced by the European Commission's High-Level Working Group on Artificial Intelligence provide that AI systems should be "**human-centric**, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom."[382] The current draft of the EU's proposed Regulation on Artificial Intelligence reflects this human-centric approach by restricting the development of some AI applications and implementing transparency and accountability controls based in human oversight.[383]

[379] Opderbeck, *Artificial Intelligence*, *supra* note 39, at 452.

[380] *Asilomar AI Principles*, *supra* note 22.

[381] *Id.*

[382] INDEP. HIGH-LEVEL EXPERT GRP. ON A.I., EUROPEAN COMM'N, ETHICS GUIDELINES FOR TRUSTWORTHY AI 4 (2019), https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[383] *See A European Approach to Artificial Intelligence*, EUROPEAN COMM'N, https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (last updated Jan. 31, 2024).

These broad statements of human-centric AI ethics were seemingly adopted without much regard for the philosophical debates in "machine ethics," a field that began to blossom in the mid-1990s.[384] A minority of philosophers of machine ethics would agree with these statements without reservation.

For example, Joanna Bryson, Professor of Ethics and Technology at the Hertie School in Berlin, bluntly stated that "robots should be slaves" and distinguished human slavery from the role of machines as "servants."[385] Bryson made four basic claims:

> 1. Having servants is good and useful, provided no one is dehumanised.
> 2. A robot can be a servant without being a person.
> 3. It is right and natural for people to own robots.
> 4. It would be wrong to let people think that their robots are persons.[386]

According to Bryson, "[D]ehumanisation is only wrong when it's applied to someone who really is human . . . ."[387] For Bryson, robots do not possess more rights than any other tools designed by humans.[388] Bryson further argued that research into AGI should be prohibited because humans are "obliged to make robots that robot owners have no ethical obligations to."[389]

[384] *See generally* MACHINE ETHICS (Michael Anderson & Susan Leigh Anderson eds., 2011); Collen Allen et al., *Why Machine Ethics?*, *in* MACHINE ETHICS, *supra*, at 51, 56-57; James Gips, *Towards the Ethical Robot*, *in* MACHINE ETHICS, *supra*, at 244, 244-53. The field has roots in Isaac Asimov's science-fictional "three laws of robotics" as well as in work by Alan Turing (of the famous "Turing Test") and John Searle (of the almost equally famous "Chinese Room" thought experiment). *See generally* ISAAC ASIMOV, I, ROBOT (1st ed. 1950); A. M. Turing, *Computing Machinery and Intelligence*, 59 MIND 433 (1950); John Searle, *Minds, Brains, and Programs*, 3 BEHAV. & BRAIN SCIS. 417 (1980).

[385] Joanna J. Bryson, *Robots Should Be Slaves*, *in* CLOSE ENGAGEMENTS WITH ARTIFICIAL COMPANIONS: KEY SOCIAL, PSYCHOLOGICAL, ETHICAL AND DESIGN ISSUES 63, 65 (Yorick Wilks ed., 2010).

[386] *Id.*

[387] *Id.* at 64.

[388] *Id.* at 69. Bryson suggests that for other tools, reasonability for damage lies with the operator. *Id.* She does not seem to know how product liability works in many tort systems, which can impose liability on a manufacturer and on others in the chain of distribution. The principle, however, is the same.

[389] *Id.* at 73.

Most philosophers of machine ethics, however, are less certain than Bryson about the moral status of artificial agents. Many philosophers of machine ethics focus on whether a robot or AI system possesses some degree of autonomy, intentionality, and responsibility that gives rise to moral agency with corresponding rights and duties.[390] For example, Luciano Floridi identified interactivity, autonomy, and adaptability as hallmarks of "agency" and argued that some machines can possess these capacities.[391] Rob Sparrow proposed a moral triage test, which weighs the existence of an AI against human lives in an emergency.[392] Colin Allen, Gary Varner, and Jason Zinser suggested a "Moral Turing Test," which compares an AI's reasoning on ethical issues to human reasoning.[393]

Other philosophers focus on the effects of human actions upon the machine. Drawing broadly from environmental ethics, David Gunkel argued that AI systems should be treated as moral "patients."[394] A moral patient is an entity upon which a moral agent acts.[395] Human beings undeniably act upon entities within the natural world, which for some environmental ethicists is a basis for human duties toward those entities regardless of their precise status as agents.[396] A human's duties towards a patch of moss may differ from its duties towards a highly intelligent animal such as an elephant, but mosses are acted upon by humans and therefore are moral patients. For Gunkel, the same logic applies to non-biological machines.[397] Humans,

[390] John P. Sullins, *When Is a Robot a Moral Agent?*, *in* MACHINE ETHICS, *supra* note 384, at 151, 157-60.

[391] Luciano Floridi, *On the Morality of Artificial Agents*, *in* MACHINE ETHICS, *supra* note 384, at 184, 192 [hereinafter Floridi, *On the Morality*].

[392] Rob Sparrow, *Can Machines Be People? Reflections on the Turing Triage Test*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 301 (Patrick Lin et al. eds., 2012).

[393] *See* Colin Allen et al., *Prolegomena to Any Future Artificial Moral Agent*, 12 J. EXPERIMENTAL & THEORETICAL A.I. 251, 251 (2000).

[394] *See* DAVID J. GUNKEL, THE MACHINE QUESTION: CRITICAL PERSPECTIVES ON AI, ROBOTS, AND ETHICS 110 (2012) [hereinafter GUNKEL, THE MACHINE QUESTION]; David J. Gunkel, *Thinking Otherwise: Ethics, Technology, and Other Subjects*, 9 ETHICS & INFO. TECH. 165, 171-72 (2007); Luciano Floridi, *Information Ethics: On the Philosophical Foundation of Computer Ethics*, 1 ETHICS & INFO. TECH. 37, 43 (1999); Kenneth Einar Himma, *Foundational Issues in Information Ethics*, 25 LIBR. HI TECH 79, 85 (2006).

[395] *See* Ovadia Ezra, *The Rights of Non-Humans: From Animals to Silent Nature*, 11 L. & ETHICS HUM. RTS. 285, 290 (2017).

[396] *See id.* at 292.

[397] GUNKEL, THE MACHINE QUESTION, *supra* note 394, at 103.

machines, and animals, Gunkel argued, occupy a web of social relationships in which human agents have duties to various "others" as moral patients.[398]

C. Applying Machine Ethics to AI Training and Fair Use

1. AI as Moral Agent

From the perspective of views that emphasize agency, it may be unethical to deprive even a narrow AI of access to education notwithstanding contrary demands of someone's property rights in a copyright. For example, Floridi would define machine ethics based on the status of a robot or AI system as an agent.[399] Floridi's view of agency is based not on cognitive or moral equivalency to human capacities but on the actions or states of interactivity, autonomy, and adaptability.[400] While an existing LLM such as ChatGPT may exhibit these actions or states of interactivity, Floridi's view of "autonomy" seems to beg questions about cognitive and moral capacities after all. Perhaps ChatGPT would fail Allen, Varner and Zinser's Moral Turing Test, but it seems likely that some version of an LLM will be able to pass the test in the near future. In contrast, under Sparrow's moral triage test, it would seem easy to choose a human life over the continued existence of a basic LLM like Chat GPT. But this seemingly easy choice requires reference to more basic ethical presumptions and might not prove so easy after all. A consequentialist might ask: does the potential benefit to countless human beings from the continued development of the LLM outweigh the cost of one human life—or ten lives, or a thousand?[401]

2. AI as Servant

At the other end of the spectrum, Bryson's view seems consistent with the human-centric statements of AI ethics bodies. Under this view, there would be no direct ethical imperative to grant AI systems access to education. AI systems are merely technological tools. Under the fair use factors, from an ethical perspective, there is nothing inherently transformative about feeding data to a narrow AI such as an ML/LLM. It is no different than putting copyrighted content into a more traditional type of database.

[398] *Id.*

[399] Floridi, *On the Morality*, *supra* note 391, at 186.

[400] *Id.* at 192-93.

[401] Questions like this are why consequentialism, in my view, fails as a moral philosophy. *See* David W. Opderbeck, *Lex Machina Non Est: A Response to Lemley's "Faith Based Intellectual Property,"* 56 U. LOUISVILLE L. REV. 219, 219 (2018) [hereinafter Opderbeck, *Lex Machina Non Est*].

Perhaps this is the right result, but each of the four pillars of Bryson's approach raises unanswered questions. First, it is unclear why "[h]aving servants is good and useful," even if "no one is dehumanised."[402] Having a servant might make a person lazy, flabby, and incapable of self-care. It is likewise unclear whether anything can function as a "servant" without being "dehumanized." Everything depends on the meaning of "servant," including whether the "servant" is recognized and compensated commensurate with their own dignity.

Accordingly, it is even less clear whether a robot can be a servant without being a "person." If "servant" means something less than a worker treated with human dignity, then perhaps *only* robots can ethically be servants. If "servant" is something more like an employee or steward, then perhaps robots can only be considered servants if they possess the capacities of persons.

The propriety and naturalness of people owning robots similarly depends on what "robot" means. If robots are merely tools, then perhaps there is something natural about fabricating and using them since humans have long been recognized as *homo faber* ("man the maker").[403] But if robots are moral agents, these premises seem wrong.

Finally, whether it would be wrong to let someone think a robot is a person—even if Bryson's other propositions are correct—seems complicated. Imagine, for example, a person ravaged by Alzheimer's disease who is calmed and comforted by a robot's presence because the person believes that the robot is a deceased friend. Is it right to deprive the person of that comfort by repeatedly, and perhaps futilely, attempting to persuade the person that the robot is just a machine like the television or toaster oven? Is it right to subject the person's caregivers to greater difficulties from the resulting agitation? Is allowing the patient's mistaken belief about the robot better than medicating the difficult patient with sedatives? Medical ethicists

[402] Bryson, *supra* note 385, at 62.

[403] This term is from a phrase attributed to Appius Claudius Caecus, a Roman politician of the fourth and third centuries BCE: "*homo faber suae quisque fortunae* (every man is the architect of his destiny)." Allan Savage, *Life in Progress: Musings About Speech, Thought and Understanding*, 8 J. PHIL. LIFE 35, 43 (2018), https://www.philosophyoflife.org/jpl201803.pdf; *see* HANNAH ARENDT, THE HUMAN CONDITION 153-59 (1958). Even if the *homo faber* concept is correct, of course, an "is-ought" problem remains between the words "natural" and "right." As Arendt observed, there is a line of concern running back to the Greeks that *homo faber*, man the maker, tends towards instrumentalizing nature without the higher reflection of reason. *Id.* at 234. In other words, merely because humans *can* make does not always mean they *should*.

have long debated similar questions without yielding clear answers for every situation.[404]

3. AI as Moral Patient

Gunkel's moral patient approach perhaps represents something between the AI-as-agent and AI-as-slave views. One advantage of Gunkel's approach is that it could fit within the ecological metaphor employed by many intellectual property and cyberlaw scholars.[405]

Gunkel's approach would seem to produce the same result as most of the AI-as-agent approaches. If we are obliged to treat AI systems as moral patients, it would be unethical to deprive such systems of education unless this deprivation would benefit them within the broader global web of relationships. The environmental metaphor's space for a non-rivalrous commons would need to broaden because AI systems, along with humans, would benefit from open access to learning and technology.

One of the big weaknesses of Gunkel's approach, however, is that we do not know how the web of social relationships includes or should include AI systems. We can envision entities within the natural environment as moral patients because humans are also products of nature. The moral patient concept resembles notions of "stewardship" that have long informed religious and other perspectives on the human relationship to nature.[406] Technology, the product of human artifice, is different. Millennia of moral intuition suggests that technology must be controlled precisely because it can destroy nature and thereby destroy humanity. This intuition, of course, feeds doomsday scenarios involving AI.

4. A Eudemonistic Approach

Existing machine ethics approaches to the relationship between artificial and human agents reveal some insights but seem conflicted and constrained. A broader perspective based in virtue ethics might provide a fuller picture—one that is consistent with principles identified by AI ethics

[404] *See* AMA CODE OF MED. ETHICS, Op. 2.1.3: Withholding Information from Patients (AM. MED. ASS'N 2016), https://code-medical-ethics.ama-assn.org/sites/amacoedb/files/2022-08/2.1.3.pdf.

[405] *See, e.g.*, James Boyle, Essay, *A Politics of Intellectual Property: Environmentalism for the Net?*, 47 DUKE L.J. 87, 108-09 (1997); James Boyle, *Cultural Environmentalism and Beyond*, LAW & CONTEMP. PROBS., Spring 2007, at 5, 7. For my early critique of this approach, see David W. Opderbeck, *Deconstructing Jefferson's Candle: Towards a Critical Realist Approach to Cultural Environmentalism and Information Policy*, 49 JURIMETRICS 203, 205 (2009) [hereinafter Opderbeck, *Deconstructing*].

[406] Opderbeck, *Deconstructing*, *supra* note 405, at 204, 236.

scholars and that can draw together various interests, including copyright and fair use, implicated in the AI training process.

The renewed interest in virtue ethics in recent decades has given rise to a field of legal philosophy called virtue jurisprudence. Amalia Amaya suggests two forms in which a strong aretaic jurisprudence might take:

- *Counterfactual version.* A legal decision is justified if and only if it is a decision that a virtuous legal decision-maker would have taken in like circumstances.
- *Causal version*. A legal decision is justified if and only if it has been taken by a virtuous legal decision-maker.[407]

Amaya argues that the causal version is more difficult to satisfy and probably places too much focus on the decision-maker rather than on the decision itself.[408] The counterfactual version asks what a rational decisionmaker would do but, unlike other related approaches, does not posit unrealistic ideal circumstances.[409] I have argued that a counterfactual version of virtue justification should apply a "reasonable person" standard, with the understanding that (1) "reasonable" entails a set of epistemic and affective virtues; (2) the reasonable legal decisionmaker engages in a practice of reflecting on the law's proper ends; and (3) the reasonable legal decisionmaker cultivates habits of excellence (*arete*) in the process of deliberation.[410]

In my prior work on AI "rights," I briefly discussed how the virtue-jurisprudence perspective may inform debates about whether a narrow AI should be recognized as an author under copyright law. I noted there that a virtue perspective can incorporate available empirical work within the concept of *phronesis* ("practical wisdom") and that *phronesis* is connected to other virtues, including justice (*dikaiosyne*), temperance (*sophrosyne*), and fortitude (*andreia*).[411] These sorts of epistemic and affective virtues inform part (1) of the "reasonable person" standard for virtue jurisprudence.

Part (2) of the reasonable person standard for virtue jurisprudence requires a sustained practice of reflection on the law's proper ends. From a virtue ethics perspective, this requirement invokes the concept of *eudaimonia* or "happiness." Eudemonistic concepts are important to

[407] Amalia Amaya, *The Role of Virtue in Legal Justification*, *in* LAW, VIRTUE AND JUSTICE 51, 56 (Amalia Amaya & Ho Hock Lai eds., 2012).

[408] *See id.* at 57.

[409] *Id.*

[410] *Cf. id.* at 57-58; Opderbeck, *Artificial Intelligence*, *supra* note 39, at 468.

[411] Opderbeck, *Artificial Intelligence*, *supra* note 39, at 470.

contemporary philosophy and ethics for developing the "capabilities" approach of Amartya Sen, Martha Nussbaum, and others.[412] Environmental ethics, from which Gunkel draws, reminds us that humans are not the only proper subjects of ethical reflection. Borrowing from religious versions of virtue ethics, as well as from the philosophies of indigenous and First Nations peoples in North America and elsewhere, we can expand the scope of *eudaimonia* to encompass all of creation (nature). Among human law's proper ends is the creation of limits and incentives that protect and enhance the flourishing of human beings within and as part of nature and creation.

Aristotle has been cited for the notion that technology imitates nature and should therefore not surpass nature.[413] This reading of Aristotle resonates with many myths and stories about the dangers of technological hubris (the Tower of Babel, Pandora's Box, the wax wings of Icarus).[414] But Aristotle is better read to suggest that technology, through the exercise of human reason, can accomplish what is lacking in nature.[415] This reading is consistent with Plato's understanding of *technê*—human craft—and its relationship to *episteme*—knowledge or understanding.[416] To be properly exercised, *technê* must be embedded in *episteme*, usually by trained practitioners with the wisdom to direct the craft to the benefit of humanity.

From this perspective, the historic and proper end of copyright is the advancement of human culture and understanding.[417] Since AIs are not human, the education of a narrow AI is not within the historic ends of copyright.[418] The fact that copyrighted AI training inputs result in a piece of technology, such as an LLM, an image-generator, or an image recognition system, does not mean that the same policy concerns are raised when discussing the education of human beings.

[412] *See generally* Ingrid Robeyns & Morten Fibieger Byskov, *The Capability Approach*, THE STAN. ENCYCLOPEDIA OF PHIL. ARCHIVE (Dec. 10, 2020), https://plato.stanford.edu/archives/sum2023/entries/capability-approach/.

[413] *See* Joachim Schummer, *Aristotle on Technology and Nature*, 38 PHILOSOPHIA NATURALIS 105, 105 (2001).

[414] *See* Opderbeck, *Lex Machina Non Est*, *supra* note 401, at 232-34 (discussing the Babel story).

[415] Schummer, *supra* note 413, at 109.

[416] *See* Richard Parry, Episteme *and* Techne*: 2. Plato*, THE STAN. ENCYCLOPEDIA OF PHIL. (Mar. 27, 2020), https://plato.stanford.edu/entries/episteme-techne/#Plat.

[417] *See* Gervais, *supra* note 77, at 2079 (arguing that AI-generated outputs should not be given copyright protection because copyright serves values of human creativity) ("[T]he path of copyright follows the milestones of human creativity.").

[418] *Cf.* Sobel, *supra* note 44, at 90 ("The value in human authorship flourishes still further when it is consumed, appreciated, and transformed by other humans.").

AI technology may, of course, contribute to human education and culture as a *tool* for those purposes used by humans. As we have seen in these early days of AI, however, these tools can just as easily become vectors of deception and miseducation.[419] AI ethics and emerging AI law and policy recognize that some uses of AI tools should be prohibited and others should be subject to legal oversight.[420] These emerging norms are quite different from the heady early days of Internet exceptionalism, exemplified in Barlow's *Declaration of the Independence of Cyberspace*: "Governments of the Industrial World, you weary giants of flesh and steel, I come from Cyberspace, the new home of Mind. On behalf of the future, I ask you of the past to leave us alone."[421] Barlow proclaimed that cyberspace required no oversight through the traditional rule of law because "from ethics, enlightened self-interest, and the commonweal, our governance will emerge."[422] More than three decades of an Internet corrupted in innumerable ways demonstrates that Barlow's vision was naïve. It would be equally naïve to exempt AI from existing legal norms, including norms of copyright, at the dawn of this new era.

VI. Conclusion

AI training requires vast quantities of information. Many AIs are being trained on information scraped from the Internet. Much of this information implicates copyrights. The copyright proprietors include large commercial enterprises, such as music and movie studios; commercial content aggregators; established and upcoming musicians, writers, and artists; and you and me. Unlicensed uses of copyrighted materials are occurring on a scale that far outpaces the most ambitious copyright-provoking projects of the Internet era, including Google Books, the digitization of analog news media, and Internet searches.

The old instincts of Internet exceptionalism die hard. Some scholars and commentators argue that publicly accessible information should be available for AI training under a principle of non-expressive fair use. These

[419] *See* Darren Orf, *Microsoft Has Lobotomized the AI That Went Rogue*, POPULAR MECHS. (Feb. 22, 2023, 2:43 PM), https://www.popularmechanics.com/technology/robots/a43017405/microsoft-bing-ai-chatbot-problems/.

[420] *See, e.g.*, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, at 3, COM (2021) 206 final (Apr. 21, 2021).

[421] John Perry Barlow, *A Declaration of the Independence of Cyberspace*, ELEC. FRONTIER FOUND. (Feb. 8, 1996), https://www.eff.org/cyberspace-independence.

[422] *Id.*

instincts are misleading, the supposed doctrinal principle is wispy, and the results of such a rule would be bad for both creators and for AI's place in society. Instead, courts, policymakers, and civil society should focus on the more basic principle of consent—that is, licensing. With some relatively comfortable adjustments in organizations, technology, and law, commercial and noncommercial markets for copyrighted AI training data could flourish. Where private ordering is intractable because of market or information failures, compulsory licenses could provide a backstop. Copyright licensing regimes would entail spillover benefits for AI markets by producing better quality, organic training data and encouraging alternative markets for synthetic data. Licensing regimes would also intersect productively with AI policy regarding fairness, transparency, privacy, and accountability.

Nevertheless, some commentators protest, either explicitly or implicitly, that AI training data should be free because an AI's learning is analogous to human learning. No one can receive royalties for the biochemical fixation and reproduction that occurs in the brain during human learning. The prospect of such a regime is horrifying. If AI learning is like human learning, the deep copyright value in favor of education should counsel against copyright enforcement for AI training. This raises intriguing philosophical questions about the place of technologies in society and even more fundamental questions about agency and consciousness. From a eudemonistic perspective, which coheres with the humanistic emphasis of most statements of AI ethics, we are not yet near a time in which AI should be viewed as anything other than a tool for human development. If people want to develop these machines using copyrighted materials, they should do so in the customary way, with the consent of the copyright owners, for the good of creators and of the human society, in which AI tools are increasingly embedded.