Copyright in AI Training Data: A Human-Centered Approach

By 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 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 on" the training data.[2] Groups of authors and other content creators have filed lawsuits against OpenAI, the creator of the text generator ChatGPT, for ingesting their content without permission to train large language models.[3] In May 2023, the U.S. Copyright Office held listening session on AI and the visual arts that focused on the use of copyrighted works in training data.[4] The Federal Trade Commission issued an investigative demand to OpenAI that includes requests for information about the sources of its training data.[5] The European Union is considering rules that would require disclosure of copyrighted material used in training data.[6] Other lawsuits, 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 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 1980's saw cases involving arcade video games and personal computers; the 1990's and early 2000's, policy choices about the Internet and cases concerning peer-to-peer file sharing and the digitization of newspaper archives; the mid-2000's, litigation over the Google Books project and cable television and cloud-based DVRs; the early 2020's, disputes about operating system API's.[7]

---

[2] *See* 17 U.S.C. §§ 101, 106.

[3] See Jonah Valdez, "Sara Silverman and Other Bestselling Authors Sue ChatGPT for Copyright Infringement," Los Angeles Times, July 10, 2023, available at https://www.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, available at https://www.reuters.com/legal/lawsuit-says-openai-violated-us-authors-copyrights-train-ai-chatbot-2023-06-29/; Getty ImagesSee Class Action Complaints in Doe v. Github, Case No. 3:22-cv-06823 (N.D.Ca. November 3, 2022); Anderson v. Stability AI LTD, Case No. 3:23-cv-00201 (N.D. Ca. January 1, 2023); Getty Images (US), Inc. v. Stability AI, Inc., Case No. 1:23-cv-00135 (D. Del. February 3, 2023).

[4] *In the Matter of Copyright on Artificial Intelligence and Visual Arts Listening Session*, May 2, 2023, transcript available at https://www.copyright.gov/ai/transcripts/230502-Copyright-on-AI-and-Visual-Arts-Listening-Session-revised.pdf.

[5] Federal Trade Commission ("FTC") Civil Investigated Demand ("CID") Schedule, FTC File No. 232-3044, available at https://www.washingtonpost.com/documents/67a7081c-c770-4f05-a39e-9d02117e50e8.pdf?itid=lk_inline_manual_4, Interrogatories ¶15, Requests for Documents ¶¶7, 13.

[6] "EU Proposes New Copyright Rules for Generative AI," April 28, 2023, available at https://www.reuters.com/technology/eu-lawmakers-committee-reaches-deal-artificial-intelligence-act-2023-04-27/

[7] *See* Parts III and IV, *infra*.

Some scholars argue that the arc of intellectual property law over these past forty years bends towards fair use.[8] They envision a broad fair use domain for "non-expressive uses" to accommodate disruptive technologies.

There are some important problems with this vision. First, the arc of fair use bends in various, sometimes inscrutable ways. It is not at all clear that any sort of non-expressive use principle can be gleaned from computer-age case law. It is even less clear that such a principle would be doctrinally and practically coherent.

Second, the AI revolution is different, both in scale and in ethical concerns. In the 1990's, 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.[9] In the early 2000's, we were astonished that researchers could find digital copies of old newspaper articles on the NEXIS database rather than rummaging through microfiche.[10] By the mid-2000's, the world was awed that Google could scan over 20 million library books.[11] Today large language model (LLM) AI's such as ChatGPT consume *billions* of files for training purposes, including publicly accessible songs, newspaper articles, books – and much, much more.[12] Technologists predict that advances in storage, communications, and computing power will allow the next generations of AIs to make equally impressive leaps in scale.[13]

Much of the scholarship on intellectual property rights in AI training data so far assumes that AI presents the same doctrinal and ethical concerns as previous generations of digital era technologies.[14] Many scholars appear rooted in a prior generation's computer and Internet exceptionalism. Intellectual property rights, they suggest, are barriers on the road to greater and greater knowledge and cultural diffusion.

AI ethics scholars and policymakers are not so sanguine.[15] AI's logarithmic growth itself raises ethical questions. And beyond mere scale lurk even deeper 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 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.

---

[8] *See* Part IV, *infra*.
[9] *See* Daniel Ionescu, "Evolution of the MP3 Player," PCWorld, October 29, 2009, available at https://www.reuters.com/legal/lawsuit-says-openai-violated-us-authors-copyrights-train-ai-chatbot-2023-06-29/.https://www.pcworld.com/article/520590/evolution_of_the_mp3_player.html.
[10] *See* Part III.B., *infra*.
[11] *See* Stephen Heyman, "Google Books: A Complex and Controversial Experiment," The New York Times, October 28, 2015, available at https://www.nytimes.com/2015/10/29/arts/international/google-books-a-complex-and-controversial-experiment.html.
[12] *See* Part II, *infra*.
[13] *See* Erik Brynjolfsson and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (W.W. Norton & Co. 2014).
[14] *See* Part IV, *infra*.
[15] *See, e.g.,* The White House, *Blueprint for an AI Bill of Rights*, available at https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

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, and human alternatives, consideration, and feedback.[16] The current draft of an EU AI Regulation reflects similar concerns.[17] 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.[18] Notably, none of these policy documents suggest principles for intellectual property.[19]

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. There is no doubt that a reproduction is made of AI training data until the machine incorporates that data into its algorithmic functions.[20] This process could be considered only transitory in a way that does not infringe the reproduction right.[21] But the underlying data does, in a sense, live on 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. A licensed use of a copyrighted work, of course, is not infringement.[22] This means a copyright owner can consent to a use, 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 search – is fair use, and that fair use therefore also must be the primary basis for using AI training data. Not so: the more prosaic rationale is consent through express or implied licenses.

Consent is also central pillar of AI ethics, particularly as they 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.[23] This pillar of AI ethics demonstrates the close connection between AI ethics and privacy law – which is, in turn, grounded in basic human rights principles. Many kinds of texts and images ingested by AIs for training contain PII. This convergence between consent in copyright and in AI ethics suggests that more robust consent mechanisms for web crawling and scraping, supported by application design principles, would go

[16] *Blueprint for an AI Bill of Rights*, available at https://www.whitehouse.gov/ostp/ai-bill-of-rights/.
[17] See "EU Proposes New Copyright Rules for Generative AI," April 28, 2023, available at https://www.reuters.com/technology/eu-lawmakers-committee-reaches-deal-artificial-intelligence-act-2023-04-27/ 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. ITechLaw, *Responsible AI: A Global Policy Framework*, available at https://www.itechlaw.org/ResponsibleAI.
[18] *The Asilomar AI Principles*, available at https://futureoflife.org/open-letter/ai-principles/.
[19] ITechLaw, *Responsible AI Framework*, *supra* Note 17.
[20] *See* Part III.A., *infra*.
[21] *See* Part III.B., *infra*.
[22] *See* 17 U.S.C. § 201.
[23] *See Blueprint for an AI Bill of Rights*, available at https://www.whitehouse.gov/ostp/ai-bill-of-rights/

a long way towards addressing future concerns about AI training and copyright along with related concerns about privacy.[24]

The third intersection relates to what seems to be the principal fair use defense raised by organizations such as OpenAI: so-called non-expressive use. Many scholars and advocates assume non-expressive uses are inherently transformative.[25] They suggest 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 factual information. Open source computer projects entail communities that vet and correct the code. Scientific 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. AI training presently is a wild west. When AI models are trained from petabytes of data scraped from the Internet, 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 as an example of uses that could be fair under the "purpose and character of the use" factor.[26] 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 this 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. 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 the ability of AI machines to educate – or miseducate – humans provides fodder for a fair use analysis.

Part II of this paper briefly reviews what AI is and how it learns. Part III discusses why AI training involves the reproduction right. This involves a careful distinction, not often made in the existing literature, 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. This part of the paper explores themes in the emerging field of machine ethics regarding the rights of AI systems. Part VI concludes.

[24] *See* Part IV.C.*, infra*.
[25] *See* Part IV, *infra*.
[26] 17 U.S.C. § 107.

II. AI TRAINING, REPRODUCTION, AND CONSENT

A. WHAT IS AI AND HOW DOES IT LEARN

Early scholarship on the of artificial agents tended to blur distinctions among existing and potential types of agents.[27] Some of the important early scholarship focuses on what today we call "strong" AI or "artificial general intelligence" (AGI).[28] AGI is an artificial agent with capacities for reason and awareness that equal or exceed human capacities.[29] As far as we know, AGI does not yet exist.[30] Some researchers and philosophers think AGI is indeed possible and likely, while others believe there is something about the relationship between mind and body that makes AGI based only in machines impossible.[31]

The kinds of AI we presently encounter, and that are set to transform our lives in the near future, are forms of "weak" or "narrow" AI – or, more accurately, forms of machine learning (ML), including large language models (LLMs) such as ChatGPT.[32] 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.[33] 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.[34]

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, creating more and more neural nodes with different connections.[35] 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."[36]

---

[27] *See* Lawrence B. Solum*, Legal Personhood for Artificial Intelligences,* 70 N.C. L. Rev. 1231 (1992).
[28] *See id;* Reece Rogers, "What's AGI, and Why are AI Experts Skeptical?," Wired, April 28, 2023, available at https://www.wired.com/story/what-is-artificial-general-intelligence-agi-explained/.
[29] *Id*.
[30] *Id.*
[31] *See Artificial Intelligence, Rights, and the Virtues*, 60 Washburn L.J. 445 (2021).
[32] *See* IBM, "What is Machine Learning?," available at https://www.ibm.com/topics/machine-learning.
[33] See AWS Training, "What is a Neural Network?," available at https://aws.amazon.com/what-is/neural-network/#:~:text=A%20neural%20network%20is%20a,that%20resembles%20the%20human%20brain.
[34] *Id.*
[35] *See* statements of Curt Levy, U.S. Copyright Office Listening Session, *supra* Note 4, 35:8-11 (stating that "[t]he training model, consisting of millions or billions of weights, analogous to the synaptic connections in the human brain, retains no copies of the training examples.").
[36] *See* Pamela Samuelson, *Generative AI Meets Copyright*, 381 Science 158, 159 (July 2023); Matthew Sag, *Copyright Safety for Generative AI*, __ Houston L. Rev. __ (forthcoming), 17-22, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593; Ben Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45, 57-59 (2017).

Image recognition is a well-established and easily understandable application of this technology. Consider a digital photograph, such as the one below, of a beach:[37]

Most people could immediately identify this photo as a "beach" scene. A little experience with actual beaches, or with photos and videos of beaches, creates pattern recognition pathways in the brain.[38] 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 something like the rule of thirds – there is a high probability the scene is a "beach" and not, say, a law school classroom.[39]

An ML image recognition system can mimic this process using the pixel data in a digital photo.[40] A typical medium-quality cell phone image contains millions of individual pixels, each with values for screen location, color and intensity.[41] 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.[42] With enough training data, the system can probabilistically distinguish "beach" photos from "classroom" photos quickly and accurately.

---

[37] Photo: Licensed from Adobe Stock.
[38] *See* Salk Institute, "How the Brain Recognizes What the Eye Sees," June 8, 2017, available at https://www.salk.edu/news-release/brain-recognizes-eye-sees/#:~:text=Our%20visual%20perception%20starts%20in,edges%20in%20the%20visual%20scenes..
[39] *See id.*
[40] *See* Kinza Yasar, "Image Recognition," TechTarget, available at https://www.techtarget.com/searchenterpriseai/definition/image-recognition?Offer=abt_pubpro_AI-Insider.
[41] Digital camera sensors are measured in megapixels. A megapixel is one million pixels. Thus, a 50 megapixel cell phone camera, such as that in the current Google Pixel 7 phones, produces images of 50 million pixels. *See* SLR Lounge, "What are Megapixels and Why Do They Matter?," available at https://www.slrlounge.com/what-are-megapixels-and-do-they-matter-minute-photography/#:~:text=Megapixels%20(MP)%2C%20which%20translates,that%20make%20up%20the%20image; Google Pixel 7 Tech Specs, available at https://store.google.com/product/pixel_7_specs?hl=en-US.
[42] Yasar, "Image Recognition," *supra* Note 40.

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.[43] To do this, the crawler must at least make a temporary copy of the target page's code. Indexes created by bots can be used for various purposes including search.[44] A web scraper not only indexes information but also retrieves and stores content, such as text and images, from the target page.[45]

Web crawling and scraping tools are readily available and easy to use.[46] There are also repositories of crawled and scraped web data available to anyone, such as Common Crawl, which 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.[47] Other crawl datasets, such as LAION, include image and "aesthetic" data derived from Common Crawl data.[48] One of LAION's "Openclip" datasets, according to the LAION website, contains 5.8 billion text-and-image pairs.[49] ChatGPT, the Large Language Model (LLM) text model that has generated so much excitement and concern, was trained in part on Common Crawl data.[50] DALL-E, the image-generation tool that generated comparable buzz, was trained in part LAION data.[51]

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, partially synthetic, and synthetic data.[52] The "Argoverse" data set, for example, includes data

---

[43] *See* AI Multiple, Web Crawler: What It Is, How It Works & Applications in 2023, available at https://research.aimultiple.com/web-crawler/.
[44] *Id.*
[45] *See* AI Multiple, In Depth Guide to Web Scraping for Machine Learning in 2023, available at https://research.aimultiple.com/machine-learning-web-scraping/.
[46] *See, e.g.,* Lucene Website Crawler and Indexer, available at https://www.codeproject.com/Articles/32920/Lucene-Website-Crawler-and-Indexer; Zyte crawler, available at zyte.com; Open Solr crawler, available at https://opensolr.com/faq/view/web-crawler
[47] *See* Common Crawl, "So You're Ready to Get Started," available at https://commoncrawl.org/the-data/get-started/.; AWS Data Exchange, available at https://aws.amazon.com/marketplace/search/results?FULFILLMENT_OPTION_TYPE=DATA_EXCHANGE&CONTRACT_TYPE=OPEN_DATA_LICENSES&filters=FULFILLMENT_OPTION_TYPE%2CCONTRACT_TYPE&trk=868d8747-614e-4d4d-9fb6-fd5ac02947a8&sc_channel=el.
[48] *See* "LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets, March 31, 2022, available at https://laion.ai/blog/laion-5b/; "LAION-Aesthetics V1," available at https://laion.ai/blog/laion-aesthetics/.
[49] "Large Scale Openclip: L/14, H/14 and G/14 Trained on Laion-2B," available at https://laion.ai/blog/large-openclip/. *Cf.* Imagenet, https://image-net.org/about.php.
[50] Dennis Layton, "ChatGPT – Show Me the Data Sources," Medium, January 30, 2023, available at https://medium.com/@dlaytonj2/chatgpt-show-me-the-data-sources-11e9433d57e8
[51] *Id.*
[52] *See* Ben Sobel, "A Taxonomy of training Data: Disentangling the Mismatched Rights, Remedies, and Rationales for Restricting Machine Learning," in Reto Hilty, Jyh-An Lee, and Kung Chung Liu, eds., ARTIFICIAL INTELLIGENCE

collected by researchers at Carnegie Mellon University and Georgia Institute of Technology using a fleet of autonomous vehicles.[53] As another example, the "Unsupervised Llamas" data set, provided by the German appliance maker Bosch, includes lidar-mapped lane markers, also for training autonomous driving systems.[54]

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.[55] 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.[56] Yet others use a crowdsourcing model to obtain base data from individual contributors.[57] A crowdsourcing model can be useful, for example, to train language models on languages other than English.[58]

Still other providers specialize in "synthetic" training data.[59] This can include partially synthetic data, which creates datasets based on deidentified or otherwise modified real data (whether open or proprietary) and fully synthetic data not derived from any real dataset.[60] 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 to entry into markets creating AI products.[61] Gal and Lynskey note that 60% of AI training data will be synthetic by 2024.[62]

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 in Part II.A., *supra*, an AI does not retain complete or partial copies of its training data. 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

AND INTELLECTION PROPERTY (Oxford Univ. Press, forthcoming), at 15-26, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3677548.

[53] *See* Ming-Fang Chang, *et al.*, *Argoverse: 3D Tracking and Forecasting with Rich Maps*, arXiv:1911.02620, November 20, 2019.

[54] *See* https://unsupervised-llamas.com/llamas/.

[55] *See, e.g.,* Webz.io, available at https://webz.io/?_gl=1*1azr1wv*_ga*NTczNjM0NTQ3LjE2ODg3NDU5ODM.*_ga_PWD5DT66E0*MTY4ODc1MDYzNy4yLjEuMTY4ODc1MDYzOS4wLjAuMA...; Scale AI, available at https://scale.com/

[56] *See id.;* https://www.cogitotech.com/about-us/; https://www.anolytics.ai/; https://www.wisepl.com/; https://www.superannotate.com/

[57] *See, e.g.,* https://appen.com/what-we-do/#Sourcing; https://appen.com/blog/ai-requires-human-touch-appen-crowd-recruiting/; https://www.defined.ai/blog/introducing-defined-ai/.

[58] https://appen.com/blog/ai-requires-human-touch-appen-crowd-recruiting/.

[59] *See, e.g.,* https://www.superannotate.com/.

[60] *See* Michal S. Gal and Orla Lynskey, *Synthetic Data*, 109 Iowa L. Rev. __ (2023)(forthcoming), available at file:///C:/Users/dopde/Documents/SSRN-id4414385.pdf.

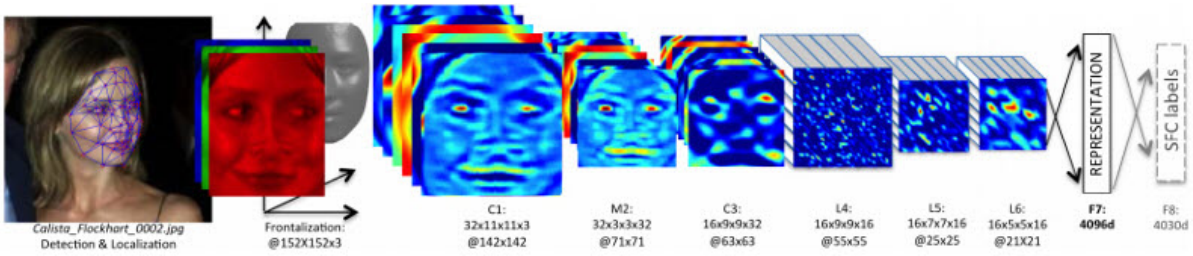
[61] *Id.* at 20, 28-29.

[62] *Id.* at 3.

training therefore might not implicate the reproduction right.[63] 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."[64]

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 comprise a reproduction. But this is not ordinarily how AI training data works.

A facial recognition AI, for example, looks at many pictures of faces and extracts mathematical relationships between various points on each face it reviews.[65] The following graphic illustrates this process:[66]



The mathematical representations labeled F7 and F8 on this graphic are stored in the system's artificial neurons.[67] As Samuelson suggests, those mathematical representations probably are not copyrightable, even aside from the fact that they are entirely machine generated.[68] They are more like facts or ideas than expression. But the original image *is* reproduced at least temporarily to generate the mathematical representations.[69] As Ben Sobel notes, this initial reproduction of the original image is a *prima facie* violation of the reproduction right.[70]

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.[71] As an

---

[63] *See* Samuelson, *Generative AI Meets Copyright*, *supra* Note 36; Sag, *Copyright Safety for Generative AI*, *supra* Note 36. Samuelson notes that, because of the idea/expression dichotomy, "[p]hotographs of cats, for instance, do not give the photographer exclusive rights to characteristic features of cats, such as their noses or facial expressions."

[64] Samuelson, *Generative AI Meets Copyright*, *supra* Note 36, at 159.

[65] *See* James Andrew Lewis and William Crumpler, *How Does Facial Recognition Work?: A Primer*, Center for Strategic and International Studies, June 2021, at 3-6.

[66] *Id.* at 6 (citing Yaniv Taigman et al., "Deepface: Closing the Gap to Human-Level Performance in Face Verification," Conference on Computer Vision and Pattern Recognition (CVPR), Facebook, June 24, 2014, https://research.fb.com/publications/deepface-closing-the-gap-to-human-level-performance-in-face-verification/).

[67] *Id.*

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

[69] *See id*.

[70] Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts at 67. In the process of training an image recognition system, reproduction of the original image might be transitory. Once the process of decomposition and extraction begins, there is no reason for the system to retain the original image. Transitory reproduction is discussed in subpart III.C., *infra*.

[71] *Id.* at 67-68.

example, Sobel argues that human facial recognition program trained only on specific portions of human portraits might not entail reproductions of the underlying portraits.[72] Perhaps Sobel is in some sense correct. Copying is a fact-specific inquiry.

But the example Sobel offers shows why *de minimus* copying is unlikely to be a good defense in most cases. That example is Labeled Faces in the Wild ("LFW"), a data set used for testing facial recognition applications.[73] LFW is maintained by the University of Massachusetts Amherst.[74] 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. . . ."[75] A review of the dataset, however, shows that it undoubtedly violates the reproduction and adaptation rights of the copyright owners in the underlying photographs, absent fair use.

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.[76] Those researchers obtained their "very large data set" of from "half a million news pictures" using a face detection tool.[77] 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):[78]

[72] *Id.*
[73] *Id.* at 67.
[74] *See* LFW Website, available at http://vis-www.cs.umass.edu/lfw/.
[75] Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts at 67.
[76] Gary B. Huang, *et al.*, *Labeled Faces in the Wild: A database for Studying Face Recognition in Unconstrained Environments*, at 9, available at http://vis-www.cs.umass.edu/lfw/lfw.pdf.
[77] Tamara L. Berg, *et al.*, *Names and Faces*, at 5-6, available at http://tamaraberg.com/papers/journal_berg.pdf.
[78] Berg, *Names and Faces*, at 22. The LFW database can be explored or downloaded at http://vis-www.cs.umass.edu/lfw/#explore. As an example, consider the entries for Brittany Spears at http://vis-www.cs.umass.edu/lfw/person/Britney_Spears.html.

Contrary to Sobel's suggestion, it seems highly unlikely that a court would find the cropped versions of these photos are non-infringing. The cropped faces still reflect decisions made by the photographers about timing, pose, lighting, expression, and effects that allow copyright in the photographs.[79] These unlicensed reproductions *prima facie* violate 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 involves much more than the mathematical tokens generated and stored in any one AI system. There are markets for AI training datasets, like LFW, that remain persistent over time.[80] 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.[81] We will return to a fuller discussion of these issues when we discuss fair use.[82] First, we discuss two reasons why this *prima facia* infringement of the reproduction right might not produce liability: transitory reproduction and consent.

B. Transitory Reproduction

As the discussion in part II.A. shows, AI training databases gleaned from crawling and scraping usually retain copies or adaptations of the original training data that likely are infringing absent consent or fair use. 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. 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 of made during training are not infringing. Transitory reproduction would not protect training databases, but it might apply to some applications that use training data.

Section 106 of the Copyright Act gives the owner of the copyright the exclusive right "to reproduce the copyrighted work in copies . . . ."[83] The Copyright 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."[84] Somewhat confusingly, under the 1976 Act, copyright is

---

[79] *See Burrow-Giles Lithographic Company v. Sarony*, 111 U.S. 53 (1884). It is possible, though not likely, that very substantial cropping could render a use *de minimus* and not infringing. *Cf. Hirsch v. CBS Broadcasting Inc.*, 2017 WL 3393845 (S.D.N.Y.)(finding copying of a photograph in a news story was not *de minimus* because "even 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.,* 2023 WL 1987206 (E.D.N.Y. 2023); *Stokes v. MilkChocolateNYC LLC*, 2023 WL 4447073 (S.D.N.Y. 2023).
[80] *See* discussion in Part __, *infra*.
[81] *See id.*
[82] Discussed in Part ___, *infra*.
[83] 17 U.S.C. § 106.
[84] 17 U.S.C. § 101.

also *acquired* when a work is "fixed in a tangible medium of expression."[85] In its definitional section the 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."[86] Fixation, then is implicated in both the acquisition of copyright and what comprises a potentially infringing copy under the reproduction right.[87]

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 1980's arcade video games set the stage for this argument.[88] These cases involved read-only memory chips (ROMs) programmed to modify or reproduce popular arcade games.[89] 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 the player's interaction.[90] Courts rejected these arguments, holding that the instructions programmed onto the original games' ROM chips sufficiently fixed the games' images, patterns, and sequences, notwithstanding variation from user input.[91]

The next phase of the debate over fixation involved computer RAM memory. 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.[92] The case involved a service company, Peak, hired to maintain and repair computers running MAI software.[93] MAI's customers were licensed to use the software, but that license did not extend to third party maintenance companies.[94] Peak argued that it never reproduced the software because it did not copy or modify any of the files on its customer's hard drives.[95] MAI argued that a copy is made every time the computer is turned on because software files are loaded into temporary RAM memory so the program can run. The court agreed with MAI.[96] This decision, which enabled

---

[85] *Id.,* § 102.
[86] *Id.*, § 101.
[87] Under the Act, *"*[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." 17 U.S.C. § 101. Aaron Perzanowski argues that, however, that "fixed" might not mean the same thing in each context. Aaron Perzanowski, *Fixing RAM Copies*, 104 N.W. U. L.R. 1067, 1088-89 (2010).
[88] *See Midway*, 547 F. Supp. 999 (N.D. Ill. 1982), *aff'd*, 704 F.2d 1099 (7th Cir. 1983), *cert. den.* 464 U.S. 823 (1983); *Stern Electronics v. Kaufmann*, 669 F.2d 852 (2d Cir. 1982); *Williams Electronics v. Arctic International*, 685 F.2d 870 (3d Cir. 1982).
[89] *See Midway Fg. Co. v. Artic Intern., Inc.*, 547 F. Supp. at 999; *Stern Electronics*, 669 F.2d at 852; *Williams Electronics*, 685 F.2d at 870.
[90] *Id.*
[91] *Id.*
[92] 991 F.2d 511 (9th Cir. 1993).
[93] *Id.* at 517.
[94] *Id.*
[95] *Id.*
[96] *Id.*

software providers to control aspects of the maintenance and repair of computer systems, attracted much scholarly and policy debate.[97]

From the video game and *MAI* cases, it seemed that the fixation with "fixed" was futile. In 2008, however, the Second Circuit in its *Cablevision* decision seemingly breathed new life into the question by holding that a temporary "buffer" copy of a video transmission was not infringing.[98] This case involved early versions of cloud-based DVRs in a time just before the streaming revolution.[99]

One of the purposes of the 1976 Copyright Act was to bring U.S. copyright law into closer harmony with the Berne Convention concerning copyright formalities and term.[100] Another purpose was to accommodate the new and growing business of cable television.[101] Cablevision was authorized to retransmit television signals to its cable television subscribers. The *Sony* case established under the fair use doctrine that individuals could record television programs on home VCRs for the purpose of "time shifting."[102] Cablevision's cloud-based DVR took a stream of its broadcast data into buffer memory that held the data for no more than 1.2 seconds.[103] This facilitated real-time rewind for users. If a customer wanted to record a program for later viewing, the data was stored in server space allocated to that customer.[104]

The Second Circuit held that the 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')."[105] The court distinguished *MAI* by noting that the "duration" requirement had not been fully litigated in that case.[106] The court concluded that the buffer copies in Cablevision's cloud DVR were not fixed under the duration requirement and therefore could not be infringing "copies."[107]

---

[97] *See* Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining is Lawful*, 53 U.S. Davis L. Rev. 893, 928-29 (2019).
[98] *Cartoon Network LP v. CSC Holdings, Inc. (Cablevision)*, 536 F.3d 121 (2d Cir. 2008).
[99] *Id.*
[100] *See* U.S. Copyright Office Copyright Timeline, "Highlight: Congress Passes the Current Copyright Act," available at https://www.copyright.gov/timeline/timeline_1950-1997.html#:~:text=The%201976%20Act%20extended%20federal,the%20author%20plus%20fifty%20years.
[101] U.S. Register of Copyrights, THE CABLE AND SATELLITE CARRIER COMPULSORY LICENSES: AN OVERVIEW AND ANALYSIS (March 1992).
[102] *Sony Corp. of America v. Universal City Studios, Inc.*, 464 U.S. 417 (1984).
[103] *Cartoon Networks*, 536 F.3d at 129-30.
[104] *Id.*
[105] *Id.* at 127.
[106] *Id.* at 128.
[107] *Id.* at 130. In *ABC v. Aereo, Inc.*, 573 U.S. 431 (2014), a similar question reached the Supreme Court. The Court concluded that Aereo's cloud DVR system comprised of an array of antennas that captured digital broadcast television, violated the "transmit cluse" because Aereo was not a cable television provider. This distinction means that *Aereo* did not address the same questions as did the Second Circuit in *Cablevision*. 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* Technavio, *Cloud DVR*

AI training might resemble the "transitory" copies of the *Cablevision* cloud DVR because, once an AI is trained on a dataset, the underlying data as such does not remain within the AI system. But the *Cablevision* court did not set any 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.[108] In fact, there is now a rapidly growing industry in data annotation.[109]

This variability 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 used by Cablevision's ingest buffer 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. 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 really was shorthand for fair use and that fair use is where an analysis of short-term copying for data processing purposes belongs.

In addition to these factual and doctrinal problems, as noted in Part II.A., from the perspective of AI policy and ethics, we do *not* want proprietors of AI systems to destroy their training data.[110] If the AI is producing undesirable results, the training data might help us understand why. Further, privacy law in many jurisdictions requires that data subjects whose personally identifiable information was used in training data have access to the data and rights of portability and rectification.[111] At the very least an AI 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.

C. CONSENT

Perhaps the most obvious and most overlooked response to copyright in AI training data gleaned from the Internet is consent. Copyright owners, of course, can "authorize" others to user their

---

Market by Platform, Type, and Geography – Forecast and Analysis 2023-2027, available at https://www.technavio.com/report/cloud-dvr-market-industry-analysis.

[108] *See, e.g.,* Vickram Singh Bisen, "Why Data Annotation is Important for Machine Learning and AI," Medium, December 21, 2019, available at https://medium.com/vsinghbisen/why-data-annotation-is-important-for-machine-learning-and-ai-5e647637c621; Cloudfactory "Data Annotation Tools for Machine Learning (Evolving Guide)," available at https://www.cloudfactory.com/data-annotation-tool-guide.

[109] *See, e.g.,* Markets and Markets, "Data Annotation and Market Component, Data Type, Application (Dataset Management, Sentiment Analysis), Annotation Type, Vertical (BFSI, IT and ITES, Healthcare and Life Sciences and Region – Global Forecast to 2027," available at https://www.marketsandmarkets.com/Market-Reports/data-annotation-and-labelling-market-20349022.html?gclid=Cj0KCQjw2qKmBhCfARIsAFy8buJF8zCNivZsjiiujvermQ2mVDyLSNBaNkjcbAFq46dngmHbNsljcA0aAkp3EALw_wcB (stating that "[t]he 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").

[110] *See, e.g.,* Blueprint for an AI Bill of Rights, *supra* Note 16, "Algorithmic Discrimination Protections" (noting that part of protection against algorithmic discrimination involves "use of representative data.").

[111] *Cf.* GDPR Art. 16.

works through assignments and licenses.[112] This is, indeed, how people in creative industries typically make money from copyrights. Of course, most copyright-protected content is not directly monetized. Many commercial websites explicitly do not make money directly from their content but exist to direct users to their products and services. Such sites typically link to a terms of service that allow users to view the content through their Internet browsers but not otherwise to make copies or distribute the content.[113] And most people who contribute online content through social media sites and the like do not make money from their content. They receive other social rewards in return for the nonexclusive licenses they give to hosting sites to publish their content.[114]

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, which would preclude other uses, including web crawling and data scraping. But these sites are routinely crawled for the purpose of Internet search without allegations of copyright infringement. In fact, web crawling is the foundation for Internet search engines, including Google. Crawling is how Google indexes the web.[115] Google, not surprisingly, never asked for anyone's permission before launching its indexing and search technology. So why aren't Google, Bing, and other search providers liable for billions upon billions of instances of copyright infringement?

The fact is that no one knows because there has never been serious test litigation over standard web search. There are some well-known early cases involving some aspects of image search, particularly the *Google v. Perfect10* and *Kelly v. Arriba Soft* cases.[116] These cases, discussed in subpart II.D below, are widely considered to establish that search is fair use. But these relatively early cases, decided by a few circuit courts, examining specific kinds of rough image-based search capabilities, seem a rickety support for the massive, globally important search business.

The more prosaic explanation is consent. In addition to their terms of use, websites by convention include a "robots.txt" file that specifies the rules for web crawlers.[117] 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. Some courts have held that a robots.txt file is a

---

[112] 17 U.S.C. §§ 106; 201(d).

[113] *See, e.g.*, jeep.com (the website for JEEP vehicles), Terms of Service, Sections 3 ("License Grant"), 4 ("Use Restrictions"), 5 ("FCA's Intellectual Property"), available at https://www.jeep.com/crossbrand_us/terms-of-use.

[114] *See, e.g.,* YouTube Terms of Service, "License to YouTube," available at https://www.youtube.com/t/terms#27dc3bf5d9.

[115] *See* Google, "How Google Search Organizes Information," available at https://www.google.com/search/howsearchworks/how-search-works/organizing-information/?utm_source=sem&utm_medium=cpc&utm_campaign=US-HSW-BKWS-PHR&utm_content=rsa&gclid=Cj0KCQjw2qKmBhCfARIsAFy8buI9oPe9OOzLrBQmVZO9VB4BXV4sY9H06_vN6wldiusUo6-7bVN6xZwaApUJEALw_wcB&gclsrc=aw.ds (stating that "[m]ost of our Search index is built through the work of software known as crawlers.").

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

[117] *See* TechTarget, "Web Crawler" definition, available at https://www.techtarget.com/whatis/definition/crawler; Google Search Central, "Introduction to robots.txt," available at https://developers.google.com/search/docs/crawling-indexing/robots/intro#:~:text=A%20robots.txt%20file%20tells,or%20password%2Dprotect%20the%20page..

technological measure under the Digital Millennium Copyright Act such that circumventing the file's restrictions is unlawful.[118] 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.[119]

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. Internet search fostered a set of norms about some kinds of crawling that facilitated search and that no one wants to test. But copyright litigation over web-scraped AI training data sets, either in conformity with robots.txt permissions or in circumvention of them, might test the current "look the other way" ethos of crawling and data mining beyond its breaking point.[120] Indeed, this seems already to be happening, 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 will argue, 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 commons and the research commons.[121] 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.[122] For example, OpenAI, the creator of ChatGPT and DALL-E, argues that its use of training data is transformative because

---

[118] *See Healthcare Advocates, Inc. v. Harding, Early, Follmer & Frailey*, 487 F. Supp. 2d 627, 643 (E.D. Pa. 2007).

[119] *See Field v. Google Inc.*, 412 F. Supp. 2d 1106 (D. Nev. 2006)(failure to configure metatags to prevent indexing constituted implied license for web crawler search indexing); *Parker v. Yahoo!, Inc.*, 88 U.S.P.Q.2d 1779 (E.D. Pa. 2008)(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, 663-66 (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 (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 robot.txt file's limitations. *hiQ Labs, Inc. v. LinkedIn Corp*., --- F. Supp. 3d ---, 2022 WL 18399982 (2022);

[120] 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 privacy. *See* Benjamin L.W. Sobel, *A New Common Law of Web Scraping*, 25 Lewis & Clark L. Rev. 147 (2021).

[121] *See, e.g.,* Carroll, *Copyright and the Progress of Science, supra* Note 97.

[122] *See* Open AI Submission, at 5 n.18; Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 Nw. L. Rev. 1607, 1638 (2009); Ben Sobel, *Artificial Intelligence's Fair use Crisis*, 41 Colum. J.L. & Arts 45, 51-57 (2017). Sobel notes that, even if there is a non-expressive use doctrine under fair use, many existing AI applications produce

"[w]orks in training corpora were meant primarily for human consumption for their standalone entertainment value" and because the outputs of the LLM are different than the training data.[123]

The 1976 Copyright 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;
>
> (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.[124]

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.'"[125]

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 is a good example. The training inputs include images of beaches. The system's output is not any kind of image at all: it is a decision upon evaluating another image. The decision – "yes that is a beach" or "no that isn't a beach" – obviously 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 some other photograph is also a beach scene. The theoretical and practical basis for this supposed doctrine, and for its application to AI training data, however, seems shaky. At the very least, it 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 the *Google Books* and *Hathi Trust* cases.[126] These cases involved scanning large volumes of books from academic and other libraries. *Hathi Trust* included books scanned by

---

expressive outputs, so the use is not really non-expressive in any event. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts at 72.

[123] Open AI Submission, at 5.

[124] *Id.*

[125] *Campbell v. Acuff-Rose*, 501 U.S. 569, 579 (1994).

[126] *The Author's Guild v. Google*, *Inc.*, 804 F.3d 202 (2nd Cir. 2015), *cert. denied*, 578 U.S. 941 (2016); *The Author's Guild v. Hathi Trust*, 755 F.3d 87 (2d Cir. 2014).

Google, the Internet Archive, and Microsoft. [127] Text-to-speech versions of copyrighted books were made available through the Hathi Trust website to individuals with visual disabilities and allowed anyone to search the text of scanned books. *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.[128]

In its Google Books decision, 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."[129] The "snippet view" did not allow users to piece together an entire book. 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."[130] The Second Circuit reached a similar conclusion about the "search" function in the *Hathi Trust* case, along with reproductions made accessible to blind persons.[131]

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 his interesting paper *Copyright for Literate Robots*, for example, 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."[132] In that paper, Grimmelmann sketches what he believes existing doctrine says, not necessarily what it should say. 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 extremely unpleasant end."[133] A similar argument, although seemingly with less hesitation about the dangers of AI is made by Professors Mark Lemley and Bryan Casey in their paper *Fair Learning*.[134]

It is not so clear, however, that existing doctrine says anything so broad about "bulk non-expressive uses." The Second Circuit's focus in *Google Books* and *Hathi Trust* was on the market for the copyrighted work, not on the degree of expression in the allegedly infringing use.[135] 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."[136] The court credited Google's and Hathi Trust's *factual* arguments that search snippets enabled by full-text scans would not erode the market for complete

[127] *Hathi Trust*, 755 F.3d 87.
[128] *Author's Guild v. Google*, 804 F.3d at 221.
[129] *Id.*
[130] *Id.* at 222.
[131] *See id*. (citing *Hathi Trust*, 755 F.3d at 98).
[132] James Grimmelmann, *Copyright for Literate Robots*, 101 Iowa L. Rev. 657, 666-67 (2016).
[133] *Id.* at 678.
[134] Mark Lemley and Bryan Casey, *Fair Learning*, 99 Tex. L. Rev. 743 (2021).
[135] *Author's Guild v. Google*, 804 F.3d at 221.
[136] *See id.*

published books.[137] This is not any kind of 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. [138] They claim that "non-expressive use" exception is "the reason most automated search and analysis tools exist in the first place."[139] The cases, however, are not so clear.

In an early image-based search case relied upon in support of a non-expressive use exception, *Perfect 10 v. Amazon*, the Ninth Circuit held that Google did not infringe the display or 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 – the "server test."[140] However, the court held that Google's use of the images was fair use.[141] Under the purpose and character of the use factor, the court concluded that "[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."[142]

Perfect 10's reasoning 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 reproduction or derivative work rights.[143] Training data is not displayed or distributed as a pointer to other information. It is copied wholesale in order to make what is arguably a derivative work.

Another early search engine case, *Kelly v. Arriba Soft*, is closer to the point.[144] Arriba Soft's crawled and scraped images to generate low-resolution image thumbnails used in its search engine.[145] 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.[146] 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.

The functional change from aesthetically pleasing work to cipher in a search engine could be similar to the change in function 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

---

[137] *See id.*
[138] *Id.* at 762.
[139] *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 mean is 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.*
[140] *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1160-63 (9th Cir. 2007).
[141] *Id.*
[142] *Id.* at 1165.
[143] *Id*. at 1159.
[144] *Kelly v. Arriba Soft*, 336 F.3d 811 (9th Cir. 2003).
[145] *Id.* at 815-16.
[146] *Id.* at 818.

are *meant for training* human beings, the transformation is not so dramatic. Moreover, in both *Perfect 10* and *Kelly*, as in most cases, 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.[147] 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.[148] The question was not so much about whether the use of the copyrighted inputs was expressive as it was about whether there was a viable market for the input works as inputs. But the transformation that might comprise a different market is not just in turning something non-expressive into something expressive. In fact, something that is at present non-expressive may have great market potential precisely because it has not (yet) been publicly expressed – that is, an unpublished work[149] There is a growing market for use of existing and new works of all kinds AI training data, which suggests that the early search engine cases may not apply. In any event, the cases focus on specific uses and markets and do not announce a generally applicable rule of non-expressive fair use.

2. Other Bases for Non-Expressive Use

In addition to the early search engine and *Google Books / Hathi Trust* cases, Mathew Sag and other scholars posit a concept of non-expressive fair use based on various snippets of copyright doctrine, including the collective work right as discussed in the Supreme Court's opinion in *New York Times v. Tasini*.[150] *Tasini* dates to an era when newspapers such as the New York Times were just

---

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

[148] Wendy Gordon, *Fair Use as Market Failure: A Structural and Economic Analysis of the Betamax Case and Its Predecessors*, 82 Colum. L. Rev. 1600 (1982).

[149] *Harper & Row v. Nation Enterprises*, 471 U.S. 539, 550 (1985)(quoting Nimmer on Copyright, § 13.05, at 13-62, n.2). As the Court in *Harper & Row* noted, "it has never been seriously disputed that 'the fact that the plaintiff's work is unpublished . . . is a factor tending to negate the defense of fair use.'" *Id*. In that case, the Court found that The Nation's unauthorized publication of portions of Gerald Ford's memoirs before their release by Harper & Row was not a fair use. *Id.*

[150] Sag *Copyright and Copy-Reliant Technology*, 103 Nw. L. Rev. at 1631; *New York Times 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 demonstrate that the right of reproduction protects only expressive substitution, meaning reproduction that is available to the public. 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. Sag *Copyright and Copy-Reliant Technology*, 103 Nw. L. Rev. at 1635-35 (citing *Davis*, 547 F. Supp. At 724 n.9; *Warner Bros., Inc. v. Am. Broad. Cos., Inc.*, 720 F.2d 231, 241 (2nd Cir. 1983); *Madrid v. Chronicle Books*, 209 F. Supp. 2d 1227, 1234 (D. Wyo. 2002); *Walker, 615* F. Supp. At 434; *Eden Toys, Inc. v. Marshall Field & Co.*, 675 F.2d 498, 501 (2nd Cir. 1982); *Durham Indus., Inc. v. Tomy Corp.*, 630 F.2d 904, 913, n. 11 (2d Cir. 1980)). In such cases, however, the issue is not wholesale copying of the underlying work, or even any literal reproduction at all. The issue in these cases is whether non-literal similarities in things like scene structure, sequence of events, and characters add up to unlawful copying. *See, e.g., Huie v. National Broadcasting Co.*, 184 F. Supp. 198, 200 (S.D.N.Y. 1960)(refusing to consider intermediate scripts and noting that "[w]e can put aside the question of slavish copying because there is no question 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

beginning to digitize their past and current print editions.[151] Text-only digital copies were made available on commercial databases such as NEXIS and on CD-ROMs.[152] Previously newspapers were archived on the analog media of microfilm or microfiche, copies of which could be obtained and indexed by libraries.[153] Plaintiffs were independent journalists who had contributed articles to publications such as the New York Times.[154] They argued 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 individually searched and viewed apart from their publication in the collective work.[155]

When *Tasini* was heard in 2001 it was viewed as a watershed moment in the developing Internet era: would collective work publishers such as the New York Times have to engage in burdensome spade work to identity decades of past writers or their heirs, and pay potentially crippling new royalties, so that the public could search and access these documents easily in digital formats? On the writer's 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 be dispersed down to individual writers?[156] Iconic American historian and filmmaker Ken Burns even weighed in with an *amicus* brief.[157]

---

screenplay" because such evidence is usually considered an unreliable measure of any non-literal similarities in the work actually alleged to infringe. *See, e.g., Walker*, 615 F. Supp. 430. In other words, the plaintiffs' claims in these cases is not that the earlier scripts and the like themselves infringed, but that the earlier versions provide 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 work 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 of non-literal infringement is how the works appear to the consuming public. 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* 1632-33 (quoting *Arnstein v. Porter*, 154 F.2d 464, 473 (2nd 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 market was the "lay public" rather than expert trained musicians because, the court condluded, "lay listeners . . . comprise the audience for whom such popular music is composed. . . ."). *Id. at* 473. 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 (2d Cir. 1992) (no fair use for sculptures based on photographs even though photographer had not yet entered sculpture market).

[151] *Id.* at 489-90.

[152] *Id.*

[153] *Id.*

[154] *Id.*

[155] *Id.*

[156] *See id.* at 504-505.

[157] *Id.* The ultimate result, as so often happens, 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. Meanwhile, Web 2.0, with its blogs and podcasts, and then the social media revolution, further disrupted every established model of journalism. [cites]

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.[158] This, the Court said, was clear from Section 201(c) of the Copyright Act, which states that

> 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.[159]

As Sag notes, the Court distinguished between collective works and databases with reference to how the content appears to an ordinary user.[160] Sag concludes that the Court thereby "reinforced that expressive communication to the public is the touchstone of copyright infringement."[161]

Tasini, however, was primarily a case about *transfers* and only secondarily about infringement. Section 201 governs transfers of rights.[162] As Justice Stevens noted in a dissent joined by Justice Breyer, the case "raise[d] an issue of first impression concerning the meaning of the word 'revision' as used in § 201(c) . . . ."[163] 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. There was no dispute that if the agreements executed by the authors 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 emphasizes the idea-expression dichotomy in favor of non-expressive use.[164] This sort of question in American copyright law has been addressed in the context of catalog and database protection, starting the *Feist* Court's treatment of the decidedly old-school technology of telephone white pages.[165] 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.[166] It is true that the idea-expression dichotomy could be

---

[158] *Id.* at 467.
[159] *Id.* (quoting 17 U.S.C. § 201(c)).
[160] Sag *Copyright and Copy-Reliant Technology*, 103 Nw. L. Rev. at 1632; *Tasini* at 501-02.
[161] Sag *Copyright and Copy-Reliant Technology*, 103 Nw. L. Rev. at 1632.
[162] *See* 17 U.S.C. § 201.
[163] *Id.* at 506 (Stevens, J. dissenting).
[164] Sag *Copyright and Copy-Reliant Technology*, 103 Nw. L. Rev. at 1631; 17 U.S.C. § 102(b).
[165] *Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340 (1991).
[166] *See Directive 96/9/EC of March 11, 1996 on the legal protection of databases,* as anticipated in Article 5 of the WIPO Copyright Treaty of 1996. In the early 2000's database protection bills were proposed in Congress but never gained substantial support. *See* Statement of David O. Carson, General Counsel, United States Copyright Office, before the Subcommittee on Courts, the Internet, and Intellectual Property Committee on the Judiciary and the Subcommittee on Commerce, Trade and Consumer Protection Committee on Energy and Commerce on the

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 arise first as to the copyrightability of the underlying works and then under the "nature of the copyrighted work" and "amount and substantiality of the portion used" fair use factors.[167] In addition, the idea-expression dichotomy also could be relevant to a claim that the mathematical tokens *resulting from* AI training are copyrightable. But as noted in Part III.C. above, the process of creating such tokens begins with reproduction of text, image, and video files and the like. In most cases, this content undoubtedly passes the low threshold of "expression" under U.S. copyright law and the entire files or substantial portions of the files are ingested.[168]

Database and Collections of Information Misappropriation Act of 2003, September 23, 2003, available at https://www.copyright.gov/docs/regstat092303.html.

[167] *See* 17 U.S.C. § 106.

[168] *Cf.* the image database examples in Part III.A., *supra*. Sag also mentions the substantial similarity test for infringement and the rejection of intermediate copying claims in the entertainment industry demonstrate that the right of reproduction protects only expressive substitution, meaning reproduction that is available to the public. These are creative but unavailing arguments. 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. Sag *Copyright and Copy-Reliant Technology*, 103 Nw. L. Rev. at 1635-35 (citing *Davis*, 547 F. Supp. At 724 n.9; *Warner Bros., Inc. v. Am. Broad. Cos., Inc.*, 720 F.2d 231, 241 (2nd Cir. 1983); *Madrid v. Chronicle Books*, 209 F. Supp. 2d 1227, 1234 (D. Wyo. 2002); *Walker, 615* F. Supp. At 434; *Eden Toys, Inc. v. Marshall Field & Co.*, 675 F.2d 498, 501 (2nd Cir. 1982); *Durham Indus., Inc. v. Tomy Corp.*, 630 F.2d 904, 913, n. 11 (2d Cir. 1980)). In such cases, however, the issue is not wholesale copying of the underlying work, or even any literal reproduction at all. The issue in these cases is whether non-literal similarities in things like scene structure, sequence of events, and characters add up to unlawful copying. *See, e.g., Huie v. National Broadcasting Co.*, 184 F. Supp. 198, 200 (S.D.N.Y. 1960)(refusing to consider intermediate scripts and noting that "[w]e can put aside the question of slavish copying because there is no question 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 non-literal similarities in the work actually alleged to infringe. *See, e.g., Walker*, 615 F. Supp. 430. In other words, the plaintiffs' claims in these cases is not that the earlier scripts and the like themselves infringed, but that the earlier versions provide 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 work 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 of non-literal infringement is how the works appear to the consuming public. 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* 1632-33 (quoting *Arnstein v. Porter*, 154 F.2d 464, 473 (2nd 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 market was the "lay public" rather than expert trained musicians because, the court condluded, "lay listeners . . . comprise the audience for whom such popular music is composed. . . ."). *Id. at* 473. 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 (2d Cir. 1992) (no fair use for sculptures based on photographs even though photographer had not yet entered sculpture market).

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

The early to mid-Internet era cases discussed above concerned aspects of digital database technologies apart from the software that makes those technologies run. Copyright protection for software as such is the large statutory elephant in the room raising its eyebrows at claims of non-expressive fair use.[169] Computer code usually is not visible to the public. In many contemporary software-as-a-service cloud applications the code remains on servers controlled by the copyright owner or its agents. Sag suggest that software is a statutory anomaly that should not dilute his broader argument.[170] But as a matter of statutory interpretation, there is no doubt that computer code is copyrightable, so the Copyright Act cannot be read to include a general fair use protection for all non-expressive uses. [171]

The Supreme Court's recent decision in *Google v. Oracle*, however, could signal more fair use latitude for at least some kinds of code inputs.[172] 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 [Android] that could be readily used by programmers."[173] The Court noted that fair use "can play an important role" in balancing statutory copyright for software against other interests in copyright law. According to the Court, as to software, fair use 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."[174]

APIs are portions of code that allow application programs to interface with a device operating system.[175] An operating system provides access to and control over a computing device's processing capabilities and hardware functions.[176] 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 or consumer can create compatible applications or devices.[177]

Java was developed as a lightweight language for applications on devices such as television set-top boxes.[178] It became widely used for web-based and desktop computer applications.[179] Google copied portions of some Java APIs without a license.[180] According to the Court, the portions copied included only "declaring code" – essentially the name of a function – and not the "task-

[169] Sag at 1638; 17 U.S.C. § 101 (definition of "computer programs"), 117.
[170] *Id.* at 1638.
[171] 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 specifying limitations on rights in computer programs is interesting, but moot. *See Google LLC v. Oracle America, Inc.*,141 S. Ct. 1183, 1197-98 (2021)(discussing history of copyright protection for computer programs).
[172] *Id.*
[173] *Id.* at 1203-04.
[174] *Id.* at 1198.
[175] *Id*.
[176] *Id.*
[177] *Id.*
[178] Abhinandan Bhatnagar, "The Complete History of Java Programming Language," Geeks for Geeks, available at https://www.geeksforgeeks.org/the-complete-history-of-java-programming-language/.
[179] *Google v. Oracle*, 151 S. Ct. at 1190.
[180] *Id.*

implementing programs" that would be called upon by the declaring code.[181] This meant that programmers familiar with Java could use well-known declaring code to implement functions in the Android operating system.[182]

This kind of use, the Court stated, "was consistent with that creative 'progress' that is the basic constitutional objective of copyright itself."[183] 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."[184] 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.[185]

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.[186] 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.

There are some surface parallels between *Google v. Oracle's* treatment of APIs and the use of copyrighted materials as AI training data, but the underlying concerns are quite different. Training data resembles APIs in that both are not themselves user applications but are necessary to facilitate user applications. But APIs are good for nothing other than serving as APIs.[187] 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 a non-expressive use because APIs have no expressive function at all.[188] The use of copyrighted material as AI training data might be a non-expressive use, but the underlying works otherwise have expressive functions.

Further, APIs are created by the developers of the operating systems, software, or devices to which they provide a programming interface.[189] If the developer can control the APIs, it can control secondary markets for systems, applications, and devices that interface with the underlying

---

[181] *Id.* at 1192-94. As Joshua Bloch and Pamela Samuelson have noted, "declaring code" is a misleading phrase. Joshua Bloch and Pamela Samuelson, *Some Misconceptions About Software in the Copyright Literature*, CSLAW'22: Proceedings of the 2nd ACM Symposium on Computer Science and the Law, November 2022, §2.2, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4198594. 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. *Id.*, §§ 2.4, 2.8. As discussed in Part ___ *supra*, however, this would not mean AI training data is non-infringing. There is no dispute that a reproduction must be made of the training data in order to produce algorithmic tokens that from an AI's "brain." Those tokens may not be copyrightable, but that is not the issue concerning training data.

[182] *Id.*

[183] *Id.* at 1203-04.

[184] *Id.* at 1207-08.

[185] *Id.* at 1208.

[186] *See U.S. v. Microsoft Corp.*, 253 F.3d 34 (D.C. Cir. 2001), *cert. den.* 122 S. Ct. 350 (2001).

[187] *See* Red Hat, "What is an API?," June 2, 2022, available at https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces.

[188] *See id.*

[189] *See id.*

product.[190] 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 as to 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, there is some support for a concept of non-expressive fair use in parts of the case law, but it is hardly a clear or well-established concept. *Google v. Oracle* adds some weight to 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.[191] Yet there are significant differences between APIs and the multifarious works that may be used as AI training data.

B. THE *WARHOL* EFFECT

The Court's most recent foray into fair use, although in a context that involves a clearly expressive use, further complicates things for any non-expressive fair use doctrine as applied to AI training data. That case involved Andy Warhol's pop art silkscreen portrait of the musician Prince based on a photograph by rock photographer Lynn Goldsmith.[192] The Warhol Foundation argued that Warhol's treatment of the photograph, in a style for which Warhol had become famous, was transformative. Writing for the majority, Justice Sotomayor 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."[193] Although transformativeness – what Justice Sotomayor called "new expression" – "may be relevant . . . it is not, without more dispositive of the first factor."[194]

Justice Sotomayor noted that the illustrative fair use purposes in section 107 offer paradigmatic examples of uses that are not merely substitutions for the underlying work.[195] But even new works that are in some sense transformative, she stated, can fall within the scope of the copyright owner's right to control derivative works.[196] This is evident, she argued, in the statutory definition of a derivative work, which includes "any other form in which a work may be recast, *transformed*, or adapted. . . ."[197] Therefore, according to Justice Sotomayor, "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

[190] *See id*.
[191] *See* 17 U.S.C. § 102(b).
[192] *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258 (2023).
[193] *Id.* at 1273.
[194] *Id.*
[195] *Id.* at 1274.
[196] *Id*. at 1275.
[197] *Id*. (quoting 17 U.S.C. § 107 (emphasis added)).

of transformation required to make 'transformative' use of an original must go beyond that required to qualify as a derivative."[198]

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.[199] If the use is commercial, this is not dispositive, but it cuts against fair use.[200] 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 incentive to create," such as parody or satire, this cuts in favor of fair use.[201] In sum, the Court stated,

> the 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.[202]

The inquiry into purpose thus is not a subjective account of the user's intent but rather is "an objective inquiry into what use was made, *i.e.*, what the user does with the original work."[203]

Applied to Warhol's treatment of the Goldsmith photo, the Court found that the original photo and Warhol's treatment served the same purpose of illustrating magazine stories about Prince and that this similarity of purpose together with the commercial nature of Warhol's use cut against fair use.[204] In response to concerns that this seemingly narrower reading of transformativeness would stifle future creativity, Justice Sotomayor 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."[205]

---

[198] *Id.* Justice Sotomayor 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 1282.
[199] *Id.* at 1276.
[200] *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, available at 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*.
[201] *Id.*
[202] *Id.* at 1277.
[203] *Id.* at 1284.
[204] *Id.* at 1279-80.
[205] *Id.* at 1286.

Warhol's limited view of transformativeness seems inconsistent with *Google v. Oracle's* more expansive view. Justice Sotomayor attempted to distinguish *Google v. Oracle* in several ways. She 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.'"[206] Because the Java code at issue in that case was "primarily functional," she suggested, it was more difficult to determine a line between unlawful copying and fair use.[207] Further, Google put the Java APIs to use in a "distinct and different computing environment," that is, in an operating system built for mobile devices rather than in desktop and laptop computers.[208] 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.[209]

Applying the *Warhol* Court's additional elements to AI training data likely will not yield predictable results. As to the first element, some AI applications are non-commercial, but many, if not most, 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 they are not really free.[210] This factor usually will cut against fair use.

As to 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. An AI application such as a text or image generator produces things that resemble traditional domains of copyright policy – text and images – but the output is generated by a machine. The Copyright Act anticipates the use of machines to fix, store, copy, distribute, and transmit copyrightable works, but it does not anticipate that machines could be responsible for copyrightable expression. Courts and commentators are only just beginning to grapple with whether AI-generated content is copyrightable, and the better answer is that it is not.[211]

Many, if not most, AI applications, however, do not produce any arguably creative output at all. The basic function of most 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 this user; and so-on. Although the copyrighted inputs used for AI training were employed for reasons well beyond the purposes of their original creation, the purpose of the training was not the purpose 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.

---

[206] *Id.* at 1274 (quoting *Google v. Oracle,* 141 S. Ct. at 1197).
[207] *See id.* at 1277 n.8.
[208] *Id.* (quoting *Google v. Oracle*, 141 S. Ct. at 1203).
[209] *Id.* 1299-1301 (Kagan, J., dissenting).
[210] *See* OpenAI Terms of Use, available at https://openai.com/policies/terms-of-use; OpenAI Privacy Policy, available at https://openai.com/policies/privacy-policy.
[211] *See* Daniel J. Gervais, *The Machine as Author*, 105 Iowa L. Rev. 2053, 2079 (2020).

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.[212] 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 to be repurposed as AI training data. When the present generation of text and image generators were trained, perhaps those markets were not yet on the horizon. But it is easy to see how such markets could be plausible, and beneficial, now that 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.[213] Oracle was not in the business of using Java to create a mobile operating system.[214] The jury also could have concluded, the Court said, that Google's use of some declaring functions for the convenience of developers did not appreciably affect existing or prospective markets for Java.[215] Further, the jury could have found that Google's Android operating system, through it incorporated some Java API declaring code, was not a market substitute for Java – that "Google's Android platform was part of a distinct (and more advanced) market than Java software."[216]

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."[217] 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.[218] The Court found that Google's copying benefitted the public because application programmers were already deeply familiar with the Java APIs.[219] Requiring programmers to learn a new set of declaring functions, the Court said, would allow Oracle to stifle innovation in new markets.[220]

---

[212] *Warhol*, 143 S. Ct. at 1276,
[213] *Google v. Oracle*, 141 S. Ct. at 1206-07. 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. *Id.* at 1195.
[214] *Id.*
[215] *Id.*
[216] *Id.* at 1207.
[217] *Id.*
[218] *Id.*
[219] *Id.* at 1208.
[220] *Id.*

The Court's discussion of quasi-antitrust lock-in effects reflects a deeper concern raised in many of the *amicus* briefs about open-source norms for APIs.[221] The proprietor of an operating system, software package, or device sometimes releases APIs publicly for free. This often happens when the underlying system, package, or device provides a kind of infrastructure for other applications. Microsoft, for example, publicly releases APIs for its Windows operating system.[222] An operating system is subject to network effects, meaning it is more valuable to each user as more users adopt the platform. 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 to all users.[223] The same is true of APIs for Apple operating systems.[224] 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. 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.[225] Alternatively, a proprietor might make APIs available for license to certain business partners, as was the case with the Java APIs at issue in *Google v. Oracle*. 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; Google had released six versions of the Android operating system without using the Java APIs at issue since the litigation commenced, accounting for more the 90% of Android devices; and Google itself had used its dominance in search to enhance Android's position in the mobile operating system market.[226] 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.[227]

2. Markets for AI Training Data and Transaction Costs

So is the vast corpora of Internet content more like declaration code in an API or is it more like an image meant to be sold for publication in a magazine? The vastness of the corpus precludes any single answer. Getty Images, for example, alleges that OpenAI used material scraped from its

---

[221] *See, e.g., Brief Amicus Curiae of the Computer & Communications Industry Association in Support of Petitioner*, available at http://www.supremecourt.gov/DocketPDF/18/18-956/89477/20190225131614070_37482%20pdf%20Band.pdf; *Brief Amicus Curiae of Developers Alliance,* available at http://www.supremecourt.gov/DocketPDF/18/18-956/89188/20190222113700973_18-956%20Amicus%20Brief%20of%20Developers%20Alliance.pdf.

[222] *See* "Build Desktop Apps for Windows," available at https://learn.microsoft.com/en-us/windows/apps/desktop/#choose-your-app-type.

[223] *See Brief Amicus Curiae of Microsoft Corporation*, available at http://www.supremecourt.gov/DocketPDF/18/18-956/89566/20190225161900311_Brief%20of%20Microsoft%20Corporation%20as%20Amicus%20Curiae.pdf.

[224] *See* "Apple Developer Documentation," available at https://developer.apple.com/documentation/.

[225] 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 https://www.teslaapi.io/*; Jamie Bailey, "How to Build a Tesla Data Dashboard with the Tesla API," *Medium*, April 15, 2020, available at https://medium.com/initial-state/how-to-build-a-tesla-data-dashboard-with-the-tesla-api-4ebee4b9827c.

[226] *Google v. Oracle*, 141 S. Ct. at 12-17-18 (Thomas, J., dissenting)(citing Case AT.40099, Google Android, July 18, 2018 (Eur. Comm'n-Competition).

[227] 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.* 1299-1301 (Kagan, J., dissenting). Justice Sotomayor's view nevertheless carried the day.

catalog – watermarks and all – to train Dall-E.[228] Getty offers its images for a fee for a variety of purposes. Markets for AI training data are only just beginning to develop, but such markets could represent a natural extension for an enterprise such as Getty with rights to millions images that are already identified and tagged.[229] It is easy to see how OpenAI's fair use arguments might fail as 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.[230] 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.[231] The same can be said for most of the photos, videos, Tik-Toks, blogs, and so-on that comprise the Internet's content layer. A fair use defense therefore seems much more robust as to this content.

As discussed in Part II.C., however, markets for AI training data are rapidly evolving.[232] Markets for the use of ordinary web content in training data are conceivable, and likely, absent a blanket fair use rule.[233] The problem for such markets is not supply or demand – it is transaction costs.[234] 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. A number of 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.

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

One set of solutions involves private ordering. As noted in Part III.A., consent – that is licensing – lies at the heart of how copyrighted materials are made available on the Internet. Rights

---

[228] cite

[229] *See* https://www.gettyimages.com/enterprise/premium-access.

[230] *See* https://www.facebook.com/groups/2341262949302876/.

[231] 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* https://www.astronomy.com/photo-submission-guidelines/.

[232] *See* Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts at 75 (stating "[d]oes training data for machine learning constitute a market that is traditional, reasonable, or likely to develop? Surprisingly, it often does.").

[233] *Cf. Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 592 (stating that "[t]he market for potential derivative uses includes only those that creators of original works would in general develop or license others to develop"); *American Geophysical Union v. Texaco, Inc.*, 60 F.3d 913, 929-30 (2nd Cir. 1994)(asking whether licensing market is "traditional, reasonable, or likely to be developed.").

[234] Sobel nevertheless suggests that some training data markets, such as for emails to be used as training data, seem "preposterous." Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts at 75. Perhaps a market for emails seems preposterous, but if so, this may be more for privacy reasons than for market reasons. A market for information made publicly available by ordinary people – social media posts, blogs, podcasts, and the like – does not seem preposterous at all, apart from transaction costs.

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.[235] Similarly, patent pool organizations bundle patent rights for core technologies such as wireless networking.[236] These solutions involve well-known trade-offs: there are still some transaction costs in the form of organizational administrative costs built into the license fees, the organizations can become forums for horizontal price agreements and other anticompetitive behavior, and the bundled content or patents might include many items of dubious value.[237] Because of competition concerns, the music performance rights societies are governed by antitrust consent decrees dating back to the 1950's and Patent pools usually must license on fair, reasonable, and non-discriminatory (FRAND) terms to avoid antitrust violations.[238]

Other 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.[239] As another example, Netflix, Amazon Prime Video, Hulu, and other streaming services aggregate film and television content and deliver it to subscribers for monthly fees.[240] Some of these services use their distribution platforms to offer sublicense subscriptions to yet other content aggregators.[241] And as yet another example, social media sites such as YouTube and TikTok aggregate content submitted by users, including both larger commercial players and individuals.

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 to the organization's customers. Of course, large commercial content aggregators such as Getty and Amazon can face criticism because they are subject to network effects and can squeeze both content contributors

---

[235] *See* www.bmi.com;/ www.sesac.com; www.ascap.com.

[236] *See, e.g.,* https://www.sisvel.com/licensing-programs/wireless-communications/wifi6/patent-pool/introduction. Patents, of course, provide different exclusive rights than copyrights – for patents, make, use, sell, or offer for sale, rather than reproduction, distribution, adaptation, and the other exclusive rights of copyright. *Cf.* 35 U.S.C. § 271; 17 U.S.C. § 106. The concept of a rights clearinghouse, however, is similar.

[237] cite

[238] cite

[239] *See* iStock by Getty Images, "Work With Us," available at https://www.istockphoto.com/workwithus.

[240] *See* Netflix.com; Hulu.com; https://www.amazon.com/gp/video/storefront?contentId=IncludedwithPrime&contentType=merch&merchId=IncludedwithPrime.

[241] *See* Amazon Prime Video subscriptions page, available at https://www.amazon.com/gp/video/storefront/ref=atv_me_pri_c_9zZ8D2_me_chn?contentType=home&contentId=store&filterId=OFFER_FILTER%3DSUBSCRIPTIONS.

and consumers. Getty, for example, has been criticized for selling licenses that include public domain content.[242]

But Getty also faces healthy market competition from large players such as Adobe and Shutterstock as well as from other small competitors.[243] Presently the stock image market is worth nearly $4 billion and is expected to grow to $7 billion over the next five years.[244] 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 Trillion by 2030.[245] Such a large markets produces 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 a variety of collective rights management organizations for AI training data involving commercially available books, music, sound recordings, television programs, and films. This could easily involve existing distributors such as Amazon, the major record labels, and established film and television streaming providers extending existing business models into licenses for AI training data. Again, network effects and market concentration are major concerns – most likely very few observers would be sanguine about Amazon dominating AI training. But the alternative seems to be equally dominant 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. The Internet's open-source ethos has always bristled at the commercialization of cyberspace. Open-source licenses, including Creative Commons and the GNU Public License, provided a mechanism through which authors could make content available for reuse under non-commercial terms.[246] Such licenses can further foster the commons through "viral" terms that require adaptations to entail similar terms.[247] In fact, Creative Commons presently is engaged in a process "to consider not only the copyright system in which CC licenses operate, but also issues of accountability, responsibility, sustainability, cultural rights, human rights, personality rights, privacy rights, data protection, and ethics."[248]

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

---

[242] *See* Mike Masnick, "Getty Images Sued Yet Again for Trying to License Public Domain Images," TechDirt, April 21, 2019, available at https://www.techdirt.com/2019/04/01/getty-images-sued-yet-again-trying-to-license-public-domain-images/.
[243] *See* Arizton, "Stock Images And Videos Market - Global Outlook & Forecast 2023-2028," June 2023, available at Stock Images And Videos Market - Global Outlook & Forecast 2023-2028.
[244] *Id.*
[245] Fortune Business Insights, "Video Streaming Market Size, Share, and COVID-19 Impact Analysis," available at https://www.fortunebusinessinsights.com/video-streaming-market-103057.
[246] *See* GNU General Public License, available at https://www.gnu.org/licenses/gpl-3.0.en.html; Creative Commons website, available at https://creativecommons.org/.
[247] *See* Creative Commons website, available at https://creativecommons.org/.
[248] Brigitte Vézina and Sarah Hinchliff Pearson, "Should CC-Licensed Content Be Used to Train AI? It Depends," March 4, 2021, available at https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/; 2023 CC Global Summit: AI & The Commons registration page, available at https://creativecommons.org/2023/06/02/2023-cc-global-summit-registration-call-for-proposals-and-scholarships-now-open/.

contributed by individuals to social media sites.[249] Individuals who contribute content through social media sites such as YouTube, TikTok, Instagram, Facebook, LinkedIn, and the like typically retain copyright and agree to terms of service regarding how the content can be used. These terms could include provisions about whether the content would be made available as AI training data. The platforms could work out some kind of revenue sharing model with users, or not, depending on how markets develop. And as organizations such as Creative Commons develop non-commercial license terms, it will become easy for individuals and organizations who wish to do so 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 non-commercial Internet content, infrastructure already exists for markets in AI training data.

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.[250] Therefore, an aggregator or social media site might not be able to protect a market for collected copyrighted training data. The 1976 Copyright Act, however, allows the divisibility of the bundle of exclusive rights under copyright.[251] If a licensor conveys to an agent the exclusive right to grant sublicenses, even if the licensor retains the right to grant other licenses, the grant is considered "exclusive" under the Copyright Act.[252] This is how stock photography providers, for example, can enforce rights against third parties.[253] But such agreements must be carefully structured to ensure that the agent / 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.[254] This would require many aggregators and

---

[249] *See* S. Dixon, "Number of Social Media Users Worldwide from 2017 to 2027," Statista, Feb. 13, 2023, available at https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (noting that there are presently 4.89 billion individual users of social media sites).

[250] *See* 17 U.S.C. § 501(b)(allowing only "legal or beneficial owner of an exclusive right under copyright" to bring an infringement action); 3 Nimmer on Copyright § 12.02[B][1]. *Minden Pictures, Inc. v. John Wiley & Sons, Inc.*, 795 F.3d 997, 1003-1006 (9th Cir. 2015).

[251] 17 U.S.C. § 201(d).

[252] *See Minden Pictures, Inc. v. John Wiley & Sons, Inc.*, 795 F.3d 997, 1003 (9th Cir. 2015). Some case law suggests that agreements must be carefully structured to ensure that the agent / 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*, 2023 WL 2482962 (N.D. Ga. 2023)(exclusive art agency agreement that did not clearly convey copyright interest insufficient for standing to sue for copyright infringement); *Greg Young Publishing, Inc. v. Zazzle, Inc.*, 2017 WL 2729584 (C.D. Cal. 2017)("exclusive representative" for negotiating licenses had no standing to sue for copyright infringement where representation agreement did not convey rights under copyrights).

[253] *Id.*

[254] See Creative Photographers, Inc. v. Julie Torres Art, LLC, 2023 WL 2482962 (N.D. Ga. 2023)(exclusive art agency agreement that did not clearly convey copyright interest insufficient for standing to sue for copyright infringement); Greg Young Publishing, Inc. v. Zazzle, Inc., 2017 WL 2729584 (C.D. Cal. 2017)("exclusive representative" for negotiating licenses had no standing to sue for copyright infringement where representation agreement did not convey rights under copyrights).

social media sites to obtain stronger copyright licenses than they presently obtain from users.[255] Other aggregators, such as Getty Images, already offer tiers in which contributors can be non-exclusive or exclusive contributors.[256]

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 30 AI startup companies deals worth over $4 billion since 2009.[257] Google might wish to consume AI training data from user content on YouTube and its other sites for free, under a claim of fair use or as a condition of its user terms of service, while selectively asserting contract or tort-based claims against others who mine data from its sites. The large entities with huge troves of user content and vested interested in AI – including the GAMAM companies – may not want a competitive market for this kind of use apart from fair use and their own contractual terms, even if they could profit from brokering the data to third parties.[258] 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, as a backstop to private ordering solutions, the Copyright Act could encode a compulsory license for AI training data. There are already, of course, numerous compulsory licenses in the law, including for sound recordings of musical works, noncommercial broadcasting, satellite retransmission, cable system retransmission, and digital audio transmission.[259] As the subject matter suggests, compulsory licensing is a common solution to copyright challenges presented by disruptive technologies that give rise to new industries.[260]

---

[255] *See, e.g.,* YouTube Terms of Service, "Licenses to YouTube," available at https://www.youtube.com/t/terms#c3e2907ca8 (stating that 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. . . .")(emphasis added).

[256] *See* Getty Images, "Work With Us," Frequently Asked Questions, "What's the Difference Between a Non-Exclusive and Exclusive Agreement," available at https://www.gettyimages.com/workwithus.

[257] Aaron Hurst, "Google Revealed to Have Acquired the Most AI Startups Since 2009," Information Age, February 18, 2020, available at https://www.information-age.com/google-revealed-acquired-most-ai-startups-since-2009-15415/.

[258] The GAMAM companies are Google, Apple, Microsoft, Amazon, and Meta.

[259] 17 U.S.C. §§ 111, 115, 118, 119, 122.

[260] For example, early cable television systems began as a hacker's project in the late 1940's, using hilltop antennas connected to coaxial cable to distribute broadcast television signals in areas with bad reception. *See* Matthew G. Anderson, "Wired: CableTV's Unlikely Beginning," Pennsylvania Center for the Book, Spring 2010, available at https://pabook.libraries.psu.edu/literary-cultural-heritage-map-pa/feature-articles/wired-cable-tvs-unlikely-beginning;Brad Adgate, "The Rise and Fall of Cable Television," Forbes, November 2, 2020, available at https://www.forbes.com/sites/bradadgate/2020/11/02/the-rise-and-fall-of-cable-television/?sh=1dfaea5e6b31. As the practice of connecting antenna "headends" to cable began to grow into an industry, FCC regulation and copyright challenges spurred by television and movie studios mounted. U.S. Register of Copyrights, THE CABLE AND SATELLITE CARRIER COMPULSORY LICENSES: AN OVERVIEW AND ANALYSIS (March 1992), 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 1909 Copyright Act. *Fortnightly Corp. v. United Artists Television, Inc.,* 392 U.S. 390 (1968). In 1974, the Court extended this holding to the reception of "distant" broadcast signals. *Teleprompter Corp. v. Columbia Broadcasting Systems, Inc.*, 415 U.S. 394 (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. THE CABLE AND SATELLITE COMPULORY LICENSES, , at ii-iv. The FCC rules allowed some broadcasters to

Compulsory licenses can be difficult to administer and are subject to criticisms that the terms quickly become outdated and unfair. Royalty calculations can become overly complex, the licensors might pay too much or the licensee might receive too little, technology may outpace the system, and the entrenched system may stifle the growth of new technologies and markets. [261] 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 copyrighted AI training data might require a protocol that is more robust than the robots.txt file both as a technological barrier to unauthorized crawling and scraping and as a permissions mechanism and accounting mechanism for authorized crawling and scraping.[262] Such a protocol would need to distinguish permitted crawling and scraping, such as for search indexing, from what is not permitted, and it would need to be difficult to circumvent.

The robots.txt protocol was, in fact, updated in 2022 – the first update since its creation in 2004.[263] As Google's own instructions make clear, however, the robots.txt file is far from foolproof.[264] 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.[265]

A more robust robots.txt-like protocol could be supported by provisions in the Copyright Act concerning "copyright management information" (CMI). The U.S. Copyright Act presently makes it unlawful to intentionally remove or alter (CMI).[266] An injured party can recover actual or

---

obtain exclusivity over some programming the cable operators were not allowed to carry *Id.*. Finally, in the 1976 Copyright Act, Congress reached a compromised that made cable television retransmission and infringement but established a compulsory licensing regime. *Id.*; 17 U.S.C. § 111.

[261] *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," Digital Music News, June 23, 2023, available at https://www.digitalmusicnews.com/2023/06/23/section-115-compulsory-license-repeal-george-johnson/; THE CABLE AND SATELLITE COMPULSORY LICENSES, at ix-xiii.

[262] *See* comments of James Gatto, Shepherd Mullin, U.S. Copyright Office Listening Session, *supra* Note 4, 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."); comments of Rebecca Blake, Graphic Artists Guild, *id.*, at 38:17-11 (stating that "[w]e see metadata and CMI as key to being able to protect artists' works in an AI environment . . . ."); J. Scott Evans, Adobe, *id,* at 40:11 to 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 the kind of proactivity among the companies that are developing this technology to give artists a tool to control their creative work."); comments of Ben Brooks, Stability AI, *id*. at 43:14 to 44:7 (stating that Stability AI supports protocols like robots.txt for consent to automated data aggregation).

[263] *See* Internet Engineering Task Force, Robots Exclusion Protocol, RFC 9309, September 12, 2022, available at https://datatracker.ietf.org/doc/rfc9309/.

[264] *See* Google, "Introduction to robots.txt," available at https://developers.google.com/search/docs/crawling-indexing/robots/intro (noting that a robots.txt file will not necessarily prevent a page from showing up in search results).

[265] *See* Zyte, "How to Avoid Web Scraping Blocks and Bans," available at https://www.zyte.com/blog/scraping-blocks-and-bans/; "How to Manage Bans and Get Data With Zyte Data API Smart Browser," available at https://www.zyte.com/blog/manage-bans-and-get-your-data-zyte-data-api/.

[266] 17 U.S.C. § 1202(b)(1).

statutory damages against a party who distributes copies of works 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.[267] As many commentators have been suggesting, the definition of CMI could be extended, either by statutory amendment or by regulation through the Register of Copyrights, to include permissions regarding use for AI training data.[268]

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.[269] 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 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, to the praise of many open Internet activists (even if John Perry Barlow still disapproved of the CDA).[270] This exemption supported countless instances of learning and creativity but also helped produce today's toxic social media culture.[271]

The Digital Millennium Copyright Act of 1998, which implemented the WIPO Copyright Treaty, included a technological anti-circumvention provision that riled open Internet and open-source advocates.[272] It also included safe harbors from secondary copyright liability for sites that took certain steps to limit infringing content.[273] Open Internet activists welcomed the safe harbors but also raised concerns that they are insufficiently attentive to fair use.[274] Like Section 230, the

---

[267] *Id.* § 1202(b)(3).

[268] *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. Comments of Rebecca Blake, Copyright Office Listening Session, *supra* Note 4, at 38:23 to 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, 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.

[269] *See, e.g.,* David R. Johnson and David Post, *Law and Borders – the Rise of Law in Cyberspace*, 48 Stan. L. Rev. 1367 (1996).

[270] 47 U.S.C. § 230; Electronic Frontier Foundation, "Section 230," available at https://www.eff.org/issues/cda230; Richard Bennett, "The Legacy of Barlow's Cyberspace Declaration of Independence," AEIdeas, February 10, 2016, available at https://www.aei.org/technology-and-innovation/telecommunications/legacy-barlows-cyberspace-declaration-independence/.

[271] *See* Michael D. Smith and Marshall Van Alstyne, *It's Time to Update Section 230*, Harvard Business Review, August 12, 2021, available at https://hbr.org/2021/08/its-time-to-update-section-230.

[272] 17 U.S.C. § 1201; Electronic Frontier Foundation, "Digital Millennium Copyright Act," available at https://www.eff.org/issues/dmca.

[273] 17 U.S.C. § 512;

[274] Electronic Frontier Foundation, "Digital Millennium Copyright Act," supra note 272.

DMCA safe harbors supported the dynamic creativity of Web 2.0 but also facilitated platform consolidation and cultures of abuse.[275]

The FCC's decision in 2002 to classify broadband cable Internet as an "information service" under the deregulatory impetus of section 706 of the Telecommunications Act of 1996 ensured a light regulatory touch rather than the more extensive regulation imposed on telecommunications services.[276] Although the Internet never became the libertarian utopia imagined by Barlow, this light touch regulation allowed its growth to be managed by technologists and community members rather than by bureaucrats.

But progressives changed their mind when the Internet backbone market became highly concentrated. Leaving cyberspace to the people turned out to mean leaving cyberspace's physical backbone to a few large corporations. It had become obvious that cyberspace is not a borderless world after all.[277] Progressives pushed for network neutrality rules in the FCC's 2015 *Open Internet Order*, although these rules were quickly reversed after the FCC's composition changed during the Trump administration.[278]

As these and many other examples suggest, the Internet is both exceptional and ordinary. Today enormous problems relating to cybercrime, surveillance, intellectual property, equal access,

---

[275] "Web 2.0" is term coined in the late 1990's for a World Wide Web that emphasizes user-generated content. *See* Kinsa Yazar, "Web 2.0," TechTarget, January 2023, available at https://www.techtarget.com/whatis/definition/Web-20-or-Web-2?Offer=abt_pubpro_AI-Insider. 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, October 13, 2020, available at https://www.jdsupra.com/legalnews/youtube-s-copyright-policy-pitfalls-23119/ (stating that "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 Technology Policy, September 27, 2016, available at 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, available at https://musically.com/2023/06/13/maria-schneiders-youtube-lawsuit-dismissed-just-before-trial/.

[276] *In the Matter of Inquiry Concerning High-Speed Access to the Internet Over Cable and Other Facilities*, FCC 02-77 (March 15, 2002). .

[277] *See* Jack Goldsmith and Tim Wu, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD (Oxford Univ. Press 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, available at 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, March 5, 2021, available at https://www.theverge.com/2021/3/5/22315224/tim-wu-net-neutrality-antitrust-big-tech-biden-administration-national-economic-council.

[278] 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. *In the Matter of Protecting and Promoting the Open Internet*, FCC 15-24 (March 12, 2015). The Open Internet Order subsequently was reversed in the 2018 "Restoring Internet Freedom Order." *In the Matter of Restoring Internet Freedom*, FCC 17-166 (January 4, 2018).

harassment, and privacy continue to bedevil cyberspace.[279] The same mix of exceptional and ordinary likely will characterize AI law and policy, but at even greater speed and scale.

In her discussion of the transformativeness fair use factor in *Warhol*, Justice Sotomayor stated that courts should 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."[280] 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. But while some AIs may serve to promote science and the arts, others may not. Indeed, it is possible that AI could severely degrade or destroy science, the arts, and other human endeavors.[281] The uncertainty combined with the scale of change counsels caution against any generalized fair use rules such as non-expressive use.[282]

As we stand 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 exceptionalism. AI industry leaders, in fact, are *asking* for regulation.[283] 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.[284]

While copyright should not drive AI policy, a copyright speedbump might entail spillover benefits for AI policy. One is privacy. Consider a parent who posts video clips of a child's birthday party on Instagram or TikTok.[285] Again, ordinarily there is also no market for these pictures; the poster perhaps wants to share with friends and family, or to get some "Likes" from the broader social

---

[279] *See, e.g.,* David W. Opderbeck, *Cybersecurity and Data Breach Harms: Theory and Reality*, 82 Maryland L. Rev. 1001 (2023).
[280] *Warhol*, 143 S. Ct. at 1282.
[281] As Professor Gary Marcus testified at a May 2023 Congressional hearing on AI safety, "[w]e have built machines that are like bulls in a china shop—powerful, reckless, and difficult to control." Testimony of Gary Marcus before Senate Judiciary Committee, Subcommittee Privacy, Technology, and the Law, May 16, 2023, available at https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-rules-for-artificial-intelligence.
[282] 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 Environmental Health Perspectives 871 (2001).
[283] *See* Courtney Rozen, "AI Leaders are Calling for More Regulation of the Tech. Here's What that May Mean in the U.S.," Washington Post, May 31, 2023, available at 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; The White House, "Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI," July 21, 2023, available at 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/#:~:text=As%20part%20of%20this%20commitment,help%20move%20toward%20safe%2C%20secure%2C.
[284] As Sobel correctly notes, "[c]ommercial 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, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts at 89.
[285] *See, e.g.,* https://www.tiktok.com/@kellee.cross/video/7220054164548832558?q=birthday%20party&t=1690232329512; https://www.instagram.com/p/CYCF2XtqUld/.

community. But these pictures also involve much broader privacy concerns. The posters chose to make clips of their children publicly available for others to view – wisely or not, and with a full understanding of the site's privacy policies and controls or not – but it seems unlikely the posters would not have wanted their children's faces used to train some else's AI. Here, restricting the use of these photos for AI training purposes through a new market mechanism could serve both the dynamic competition purposes of copyright and privacy values by giving the poster more control over how the clips are used.

A second spillover benefit may relate to data quality. ChatGPT, for example, is the uber digital native, born and raised on the net, freely accessing every dark corner of the web. Sadly, but not surprisingly, ChatCPT can become predatory, racist, antisocial, dishonest, and casually violent.[286] Because of such concerns, in March, 2023, a group of technologists, academics, and business and policy leaders issued a letter calling for a moratorium on some AI development and research.[287] Their recommendations included restrictions on access to certain kinds of computing power "[t]o prevent reckless training of the highest risk models . . . ."[288] Restricting computing power likely is an unwise and unrealistic goal under U.S. law, since the computer power in question is privately owned.[289] But restricting the consumption of the Internet's dark reaches through copyright is quite feasible.

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 license mechanism, all the bad data will still be available as well. But the *cost* of data will limit recklessness. Nobody wants to pay for bad data. A cost mechanism likely would accelerate markets for entities who 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 at far greater scale. Licensing costs thereby would internalize to AI application providers the externality costs of models produced with bad data while producing positive spillovers in these new data quality industries.

---

[286] *See, e.g.,* Kyle Wiggers, "Researchers Discover a Way to Make ChatGPT Consistently Toxic," April 12, 2023, available at https://techcrunch.com/2023/04/12/researchers-discover-a-way-to-make-chatgpt-consistently-toxic/; Prashnu Verma and Will Oremus, "ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused," The Washington Post, April 5, 2023, available at 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, December 8, 2022, available at https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/.

[287] "Pause Giant AI Experiments: An Open Letter, March 22, 2023, available at 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.

[288] *Policymaking in the Pause*, April 19, 2023, available at https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf, at 8.

[289] Prohibiting Google, Amazon, or Microsoft, for example, from using their own computing facilities could comprise a regulatory taking worth billions of taxpayer dollars. U.S. Const. amend. V, Clause 5; *Pennsylvania Coal Co. v. Mahon*, 260 U.S. 393, 415 (1922).

The effect on the market factor, then, could weigh against fair use even of non-commercial content for AI training data, particularly in light of the role copyright might play in connection with AI policy. At the same time, the question how AI may relate to *human* sciences and arts surfaces 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 basic than a novel theory of non-expressive use: does training data *educate* an AI, and 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 paper 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 & repurposing it in new works. Oops. I guess I should have been paying compulsory licenses."[290] 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,

> The 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.[291]

From the dawn of the Internet era until today, the energy around open source, open access, open data, and the information commons historically has been about human learning and development, and understandably so.[292] If training an AI is analogous to educating a human being, deeper copyright concerns apply, 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 the 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

---

[290] Emails on file with author.

[291] Comments of Curt Levy, U.S. Copyright Office Listening Session, *supra* Note 4, at 21:7-13. The Committee for Justice is a conservative think tank. *See* Committee for Justice, "About the Committee for Justice," available at https://www.committeeforjustice.org/about. Mr. 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. Mr. 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 synapse. When the AI is shown an example, the synaptic strengths or weights are slightly modified . . . and we call that learning." *Id.* at 52:18 to 53:10.

[292] *See, e.g.,* Carroll, *Copyright and the Progress of Science, supra* Note 97.

days of Anglo-American copyright, education has been recognized as a value that limits the scope of the property right. This value was realized early in the 18th Century English law's tolerance for abridgements.[293] It was a common practice at that time for publishers to produce abridged versions of lengthier works so that the underlying work's ideas could be made available to a broader public.[294] Samuel Johnson, a great literary celebrity of 18th Century England, described the purpose of abridgements in his unpublished 1739 manuscript *Considerations on the Case of Dr. T-S Sermons Abridged 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.[295]

Johnson acknowledged that abridgment might lessen the economic value of the underlying work, but asserted that this burden was outweighed by "the advantage received by mankind from the easier propagation of knowledge."[296]

The seemingly absolute exception for abridgement in early some early English copyright cases, did not directly carry over into later American copyright law.[297] In *Folsom v. Marsh*, a case at the roots of American fair use doctrine, Justice Story found that an abridgement of the complete works of George Washington infringed the original publisher's copyright.[298] Evaluating a case in equity for injunctive relief, Justice Story suggested several factors to determine whether a quotation or abridgment was infringing: "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."[299]

The *Folsom v. Marsh* factors came to inform the four fair use factors in the 1976 Copyright Act, which does emphasize that teaching, scholarship, and research are potential examples of fair use.[300] The "purpose and character of the use" factor suggests that "nonprofit educational purposes" would tip the scales towards fair use.[301] Copyright's emphasis on education also is evident in specific statutory exemptions for libraries, archives, and online teaching (the TEACH Act).[302] Of course, the library and archival exemptions and the TEACH Act are limited, and educational uses

---

[293] *Id*. at 1375-76 (2011).
[294] *Id.*
[295] Samuel Johnson, *Considerations on the Case of Dr. T-S Sermons Abridged by Mr. Cave* (1739) ¶19, in THE YALE DIGITAL EDITION OF THE WORKS OF SAMUEL JOHNSON, available at http://www.yalejohnson.com/frontend/sda_viewer?n=112220#. 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.
[296] *Id*, ¶20.
[297] *See Matthew Sag, The Prehistory of Fair Use,* 76 Brooklyn L. Rev. 1371 (2011).
[298] *Folsom v. Marsh*, 9 F. Cas. 342 (1841).
[299] *Id*. at 348; 17 U.S.C. § 107.
[300] 17 U.S.C. § 107.
[301] *Id.*
[302] 17 U.S.C. § 107, 108, 110.

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.[303]

Education remained an important consideration in the developing concept of fair use internationally. The Berne Convention of 1886, which the U.S. did not initially join, included a specific provision 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."[304] The U.S. at least partially came into compliance with the Berne Convention via the 1976 Copyright 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." 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) says "[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." The World Summit on the Information Society (WSIS) Declaration of Principles states "[w]e recognize that education, knowledge, information and communication are at the core of human progress, endeavour and well-being. . . ."[305] To this end, the WSIS Declaration 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."[306]

Nobody knew until recent decades that human learning is facilitated by physical and chemical connections among neurons. In past ages, human beings knew that the brain had something to do with this kind of cognition, but they might have assigned higher levels of understanding to the soul, spirit, or mind as a kind of non-material property. Even today, philosophers, scientists, and theologians debate over whether such higher levels of cognition really can be reduced entirely to the material structure and chemistry of the brain.[307] Indeed, this is the essential question in the debate over whether AGI really is possible. Human cognition may be ineluctably tied to a human body. Or 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 either event, we now know that human learning from copyrighted materials involves biochemical reproduction,

---

[303] *See Hachette Book Group, Inc. v. Internet Archive*, --- F. Supp.3d ---, 2023 WL 2623787 (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 *7.

[304] BERNE CONVENTION FOR THE PROTECTION OF LITERARY AND ARTISTIC WORKS, September 9, 1886 ("Berne Convention"), Art. 10(2).

[305] ¶8.

[306] ¶26.

[307] For my contribution to this debate, *see* David W. Opderbeck, THE END OF THE LAW? THEOLOGY, NEUROSCIENCE, AND THE SOUL (Wipf & Stock / Cascade 2021).

including transitory copies as information is ingested, longer term copies of things committed to memory, and the storage of chemical algorithmic tokens representing patterns and decision points.

Perhaps the exclusion of these biochemical functions from the definition of reproduction in copyright law is based on a faulty, pre-scientific philosophy of mind. In 1908, in *White-Smith Music Pub. Co. v. Apollo Co.*, the Supreme Court held that a player-piano roll was not a copy of the music inscribed on it because music is perceived by the ear.[308] In the 1909 Copyright Act, 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.[309] If we theoretically 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 lines of computer source code, maybe 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 books that are otherwise lawfully reproduced and distributed. The biochemical functions of the human brain as a limit on copyright runs deeper than a merely 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.

B. AI Ethics and Three Perspectives on Machine Ethics

Some statements of AI ethics suggest that AIs cannot possess anything like human rights. Instead, AIs are tools that serve humans.[310] The highly regarded Asilomar AI Principles, for example, state that "[t]he goal of AI research should be to create not undirected intelligence, but beneficial intelligence" (meaning beneficial for humans).[311] 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.
>
> . . . .

[308] 209 U.S. 1 (1908).
[309] 17 U.S.C. § 115.
[310] Portions of this section are drawn from my paper *Artificial Intelligence, Rights, and the Virtues*, 60 Washburn L.J. 445 (2021).
[311] The Future of Life Institute, Asilomar AI Principles, available at https://futureoflife.org/ai-principles/, Principle 1 (Research Goal).

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.[312]

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 state 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."[313] 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.[314]

These broad statements of human-centric AI ethics seem to have been adopted without much regard for philosophical debates in "machine ethics," a field that began to blossom starting in the mid-1990's.[315] A minority of philosophers of machine ethics would agree with these statements without reservation.

Joanna Bryson, for example, is particularly blunt in stating that "robots should be slaves."[316] Bryson distinguishes human slavery from the role of machines as "servants." Bryson's makes four basic claims:

1. Having servants is good and useful, provided no one is dehumanized.
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.[317]

According to Bryson, "dehumanization is only wrong when it's applied to something that really is human. . . ."[318] For Bryson, this means robots have no more rights than any other tools designed

---

[312] *Id.*

[313] European Commission, Independent High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, April 8, 2019, available at file:///H:/Downloads/AIHLEG_EthicsGuidelinesforTrustworthyAI-ENpdf.pdf, at 4. (emphasis in original).

[314] citesd

[315] *See generally* Michael Anderson and Susan Leigh Anderson, eds., MACHINE ETHICS (Cambridge: Cambridge Univ. Pres 2011); Collen Allen, Wendell Wallach, and Iva Smit, "Why Machine Ethics?," in MACHINE ETHICS, at 56-57; James Gips, "Towards the Ethical Robot," in MACHINE ETHICS, at 244-253. 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* Isaac Asimov, I, ROBOT (New York: Gnome Press 1st ed. 1950); Alan Turing, *Computing Machinery and Intelligence*, LIX (236) Mind 433 (1950); John Searle, *Minds, Brains, and Programs*, 3 Behavioral and Brain Sciences 417 (1980).

[316] Joanna J. Bryson, "Robots Should be Slaves," in Yorick Wilks, ed., Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues (Amsterdam: John Benjamins Publishing Co. 2010), at 63-74.

[317] *Id.* at 65.

[318] *Id*. at 64.

by humans.[319] Bryson further argues that research into AGI should be prohibited, because humans are "obliged to make robots that robot owners have no ethical obligations to."[320]

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 giving rise to moral agency with corresponding rights and duties. There are many views within this family of positions. For example, Luciano Floridi identifies interactivity, autonomy, and adaptability as hallmarks of "agency" and argues that some machines can possess these capacities; Rob Sparrow proposes a moral triage test, in which the existence of an AI is weighted against human lives in an emergency; and Colin Allen, Gary Varner, and Jason Zinser suggest a "Moral Turing Test," which would compare an AI's reasoning on ethical issues to human reasoning.[321]

Yet other philosophers shift the focus away from the agency of the machine to the effects of human action upon the machine. Drawing broadly from environmental ethics, David Gunkel argues that AI systems should be treated as moral "patients."[322] A moral patient is an entity upon which a moral agent acts. Human beings undeniably act upon entities within the natural world, which for some environmental ethicists is a basis for human duties toward entities within the natural environment regardless of their precise status as agents. 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. Humans, machines, and animals, Gunkel argues, occupy a web of social relationships in which human agents have duties to all these various "others" as moral patients.[323]

C. Applying Machine Ethics to AI Training and Fair Use

1. AI as Moral Agent

Within this family of views emphasizing agency, it could be unethical to deprive even a narrow AI of access to education notwithstanding contrary demands of someone's property rights in a copyright. Consider first the views of Floridi and others who would base machine ethics on the status of a robot or AI system as an agent. It is important to remember that Floridi's view of agency

---

[319] *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.
[320] *Id.* at 73.
[321] *See See* John P. Sullins, "When is a Robot a Moral Agent?," in Machine Ethics, at 151-161; Luciano Floridi, "On the Morality of Artificial Agents," in Machine Ethics, supra Note 30, at 184-211; Rob Sparrow, "Can Machines Be People? Reflections on the Turing Triage Test," in Patrick Lin, Keith Abney, and George A. Bekey, eds., The Ethical and Social Implications of Robotics (Cambridge, MA: MIT Press 2011), 301-316; Colin Allen, Gary Varner, and Jason Zinser, *Prolegomena to Any Future Artificial Moral Agent*, 12 J. Expt. Theor. Artif. Intell. 251 (2000).
[322] David J. Gunkel, The Machine Question: Critical Perspectives on AI, Robots, and Ethics (Cambridge, MA: MIT Press 2012); Gunkel, *Thinking Otherwise: Ethics, Technology, and Other Subjects*, 9 Ethics and Information Technology 165 (2007); Luciano Floridi, *Information Ethics: On the Philosophical Foundation of Computer Ethics*, 1 Ethics and Information Technology 37 (1999); Kenneth Einar Himma, *Foundational Issues in Information Ethics*, 25 Issues in Information Ethics 79 (2006).
[323] Gunkel, The Machine Question, Chapter 3.

is based not on cognitive or moral equivalence to human capacities but on the actions or states of interactivity, autonomy, and adaptability. An existing LLM such as ChatGPT might already exhibit these actions or states, although Floridi's category of "autonomy" seems to beg questions about cognitive and moral capacities after all. Perhaps ChatGPT could not yet pass Allen, Varner and Zinser's Moral Turing Test, but it seems likely that some version of an LLM will be able to do so in the near future. Under Sparrow's moral triage test, in contrast, 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 take pause: does the potential benefit to multitudes of humanity from the continued development of the LLM outweigh the cost of one human life – or ten lives, or a thousand?[324]

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.

Perhaps this is the right result, but each of the four pillars of Bryson's approach raises unanswered questions. First, it is unclear whether, why, or when "[h]aving servants is good and useful" even if "no one is dehumanized." Having a servant might make a person lazy, flabby, and incapable of self-care. It is likewise unclear whether anyone 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.

This means 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*.[325] 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

[324] Questions like this are why consequentialism, in my view, fails as a moral philosophy. *See* Opderbeck, *Lex Machina Non Est: A Response to Mark Lemley's 'Faith Based Intellectual Property,'* 56 Louisville L. Rev. 219 (2017).

[325] Attributed to Appius Claudius Caecus: "*Homo faber suae quisque fortunae." See* Hannah Arendt, THE HUMAN CONDITION (Univ. Chicago Press 1958). Even if the *homo faber* concept is correct, of course, an "is-ought" problem remains between the words "natural" and "right."

Alzheimer's disease who is calmed and comforted by the presence of a robot the patient believes is a deceased friend. Is it right to deprive the person of that comfort by repeatedly, and perhaps futilely, attempting to persuade the patient that the robot is just a machine like the television or toaster oven? Is it right to subject the patient's caregivers to greater difficulties from an agitated patient? Is allowing the patient's mistaken belief about the robot better than medicating the difficult patient with sedatives? Medical ethicists have long debated similar questions without yielding clear answers for every situation.[326]

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.[327]

This approach would seem to produce the same result as most of the AI-as-agent views. 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 yet know how the web of social relationships does or should include AI systems. We can envision entities within the natural environment as moral patients because we 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. 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 scholars and that can draw together various interests, including copyright and fair use, implicated in the AI training process.

---

[326] *See* AMA Code of Ethics, "Withholding Information from Patients," available at https://code-medical-ethics.ama-assn.org/ethics-opinions/withholding-information-patients#:~:text=Except%20in%20emergency%20situations%20in,or%20consent%20is%20ethically%20unacceptable..

[327] *See, e.g.,* James Boyle, *A Politics of Intellectual Property: Environmentalism for the Net?*, 47 Duke L.J. 87 (1997); Boyle, *Cultural Environmentalism and Beyond*, 70 Law & Contemp. Probs. 5 (2007). For my early critique of this approach, see Opderbeck, *Deconstructing Jefferson's Candle: Towards a Critical Realist Approach to Cultural Environmentalism and Information Policy*, 49 Jurimetrics 203 (2009).

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 a strong aretaic jurisprudence might take: causal and counterfactual, as follows:

- *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.[328]

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.[329] The counterfactual version is asks what a rational decisionmaker would do but, unlike other related approaches, does not posit unrealistic ideal circumstances.[330] 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 decision maker engages in a practice of reflecting on the law's proper ends; and (3) the reasonable legal decision maker cultivates habits of excellence (*arete*) in the process of deliberation.[331]

In my prior work on AI "rights," I briefly discussed how this perspective might 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*). These sorts of epistemic and affective virtues inform part (1) of my "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. In virtue ethics perspective, this invokes the concept of *eudaimonia* or "happiness." Eudemonistic concepts are important to contemporary philosophy and ethics of development in the "capabilities" approach of Amartya Sen, Martha Nussbaum, and others.[332] 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 / creation.

---

[328] Amalia Amaya and& Ho Hock Lai, *Of Law, Virtue and Justice – An Introduction*, in LAW, VIRTUE AND JUSTICE 6 (Amalia Amaya & Ho Hock Lai eds., Bloomsbury 2012) at 56-57.
[329] *Id.*
[330] *Id.*
[331] *Cf. id.* at 57-58; David W. Opderbeck, *Artificial Intelligence, Rights, and the Virtues*, 60 Washburn L.J. 445 (2021).
[332] *See generally* Robeyns, Ingrid and Morten Fibieger Byskov, "The Capability Approach", *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), available at https://plato.stanford.edu/archives/sum2023/entries/capability-approach/.

Aristotle historically was cited for the notion that technology imitates nature and therefore should not surpass nature.[333] 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.[334] But Aristotle is better read to suggest that technology, through the exercise of human reason, can complete what is lacking in nature.[335] This reading is consistent with Plato's understanding of *technê* – human craft – and its relationship to *episteme* -- knowledge or understanding.[336] To be properly exercised, t*echnê* 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.[337] Since AIs are not human, the education of a narrow AI is not within the historic ends of copyright.[338] The fact that copyrighted training inputs into an AI results in a piece of technology – an LLM, an image-generator, or the like – therefore is not analogous for copyright purposes to a human learning from purchased or licensed content, nor is it in itself a "transformative" use within the ends of copyright law. "Transformative," properly understood, should refer to the effect of the end product on human beings in relation to the historic goals of copyright.

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.[339] AI ethics and emerging AI law and policy recognize that some uses of AI tools should be prohibited and that other should be subject to legal oversight.[340] These emerging norms are quite different from the heady early days of Internet exceptionalism, exemplified in John Perry 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."[341] 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

---

[333] *See* Joachim Schummer, *Aristotle on Technology and Nature*, 28 Philosophia Naturalis 105 (2001).
[334] *Cf.* David W. Opderbeck, *Lex Machina Non Est: A Response to Mark Lemley's 'Faith Based Intellectual Property,'* 56 Louisville L. Rev. 219 (2017)(discussing the Babel story).
[335] Schummer, *supra* Note 333.
[336] *See* Parry, Richard, "Episteme and Techne", The Stanford Encyclopedia of Philosophy (Winter 2021 Edition), Edward N. Zalta (ed.), available at https://plato.stanford.edu/entries/episteme-techne/#Plat.
[337] As Daniel Gervais has noted, "the path of copyright follows the milestones of human creativity." Daniel J. Gervais, *The Machine as Author*, 105 Iowa L. Rev. 2053, 2079 (2020) (arguing that AI-generated outputs should not be given copyright protection because copyright serves values of human creativity).
[338] *Cf.* Sobel, *Artificial Intelligence's Fair use Crisis*, 41 Colum. J.L. & Arts at 90 (noting that "[t]he value in human authorship flourishes still further when it is consumed, appreciated, and transformed by other humans.").
[339] *See* Darren Orf, *Microsoft Has Lobotomized the AI that Went Rogue*, Popular Mechanics, February 22, 2023, available at https://www.popularmechanics.com/technology/robots/a43017405/microsoft-bing-ai-chatbot-problems/.
[340] *See, e.g.*, European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence*, April 4, 2021, available at https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.
[341] John Perry Barlow, *Declaration of the Independence of Cyberspace*, available at https://www.eff.org/cyberspace-independence.

emerge."[342] 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 public Internet. Much of this information is subject to copyrights. The copyright proprietors include large commercial enterprises such as music and movie studios; commercial content aggregators such as Getty Images; established and upcoming musicians, writers, and artists; and you and me. This is unlicensed copyright use 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 search.

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 instincts are misleading, the supposed doctrinal principle is wispy, and the results of such a rule would be bad both for 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 non-commercial 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 also would intersect productively with AI policy regarding fairness, transparency, privacy, and accountability.

Some commentators nevertheless 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 an 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. The comparison between AI and human learning at this stage of AI technology is an anthropomorphic fallacy. 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 increasingly are embedded.

[342] *Id.*