Responsible AI Starts with Licensing

*Roy S. Kaufman**

"Many people in this world are not raised to understand the concept of consent, in all walks of life, and it's important that abusers of consent not be treated as victims when they are rightfully exposed."

- X Gonzalez[1]

"Some customers are concerned about the risk of IP infringement claims if they use the output produced by generative AI. This is understandable, given recent public inquiries by authors and artists regarding how their own work is being used in conjunction with AI models and services.

To address this customer concern, Microsoft is announcing our new Copilot Copyright Commitment. As customers ask whether they can use Microsoft's Copilot services and the output they generate without worrying about copyright claims, we are providing a straightforward answer: yes, you can, and if you are challenged on copyright grounds, we will assume responsibility for the potential legal risks involved."

- Microsoft[2]

---

* Roy S. Kaufman; Managing Director of Business Development, Managing Director of Government Relations, Copyright Clearance Center. Columbia Law School Class of 1991. This Article is a companion to a talk given at Columbia Law School on September 27, 2024.

1. *Emma Gonzalez Quotes*, BRAINYQUOTE, https://www.brainyquote.com/quotes/emma_gonzalez_1017731 (last visited Mar. 18, 2025).

2. Brad Smith & Hossein Nowbar, *Microsoft Announces New Copilot Copyright Commitment for Customers*, MICROSOFT (Sept. 7, 2023), https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/ [https://perma.cc/9YKH-9YDV] [https://web.archive.org/web/20250125001027/https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/]. On September 7, 2023, Microsoft's market cap was 2.4 trillion dollars. *See Microsoft Market Cap Sep 2023*, STATMUSE, https://www.statmuse.com/money/ask/microsoft-market-cap-sep-2023 [https://perma.cc/Q7TY-H634] [https://web.archive.org/web/20250216042111/https://www.statmuse.com/money/ask/microsoft-market-cap-sep-2023] (last visited Feb. 15, 2025).

I. RESPONSIBLE AI STARTS WITH LICENSING

Responsible AI starts with licensing. AI outcomes are strengthened by reliance on responsibly sourced, high-quality copyrighted works. Consent and human centricity are hallmarks of human advancement, and also will enable better AI.

The default position of our copyright system is that the person who wishes to use copyright protected works must seek out and obtain a license before engaging in conduct that would implicate any rights protected by copyright law to avoid an infringement claim.[3] This will only seem fair to most observers, given the intellectual and economic labor involved in creating original works, and also in light of the natural and economic justice of granting the creator the right to determine how her works will be used.

Of course, the copyright system contains exceptions in certain cases where authorization may not be needed at all (e.g., fair use in the U.S.[4]) or where rights are limited to non-negotiated licenses (e.g., non-voluntary licenses and/or levies[5]). But these exceptions or limitations to rights, in order to serve the public interest of promoting the creation and distribution of original creative works (as well as to comply with international law[6]), must be carefully circumscribed to avoid unfairly prejudicing the legitimate interests of the author. To be acceptable under the Berne 3-Step test,[7] they exist in areas of market failure.[8]

---

3. *See, e.g.*, 17 U.S.C. § 501.

4. *See* 17 U.S.C. § 107.

5. *See* INTERNATIONAL FEDERATION OF REPRODUCTION RIGHTS ORGANISATIONS, COPYRIGHT LEVIES AND REPOGRAPHY (2008); *see also What Is Statutory Licensing?*, COPYRIGHT AGENCY, https://help.copyright.com.au/hc/en-gb/articles/360000006116-What-is-statutory-licensing (last visited Apr. 6, 2025) ("Statutory licences (or statutory exclusions from infringement) allow certain uses of copyright material, without the permission of the rights holder, subject to the payment of equitable remuneration.").

6. Specifically, to comply with the Berne Three-Step Test, which provides: "It shall be a matter for legislation in the countries of the Union to permit the reproduction of such works in certain special cases, provided that such reproduction does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author." Berne Convention for the Protection of Literary and Artistic Works art. 9(2), Sept. 9, 1886, *as revised* July 24, 1971, *and amended* Sept. 28, 1979, S. TREATY DOC. NO. 99-27 (1986). Exceptions outside of the test are not Berne compliant.

7. *See id.*

8. *See* Neil Turkewitz, *Consent and Compensation: Resolving Generative AI's Copyright Crisis—A Review*, MEDIUM (May 29, 2024), https://medium.com/@nturkewitz_56674/copyright-2023-neil-turkewitz-2bf3772e0114 [https://perma.cc/Z66M-W9BD] ("[U]se of original content for training AI is a thoroughly consumptive use rather than a secondary one. Indeed, licensing one's works for AI training might represent a single transaction that exhausts the entire value of the underlying work. As such, it is vital that we eschew any mechanisms that undermine the value of that transaction — including opt-outs, non-voluntary licensing, extended collective licensing or levies. These are all tools that might be suitable for uses which are secondary in nature, but not for central economic activities. International law as expressed in the Berne Convention, the so-called WIPO Internet Treaties, and the WTO TRIPS Agreement also compels such an approach given the prohibition of formalities (opt-out would represent a formality) and the inability to create or maintain exceptions that would permit uses that conflict with a normal exploitation of works. While non-voluntary licenses and levies might mitigate some of the likely prejudice of unauthorized uses, they are strictly prohibited where they would, as would be the case with training AI, conflict with a normal exploitation of the work. In short, sound public policy and international law compel us to ensure the effective exercise of consent in an environment unobstructed by conditions extraneous to the exercise of consent.").

AI training generally requires the express and voluntarily granted consent of the author. Any other approach threatens fundamental values underlying our copyright system. This observation is grounded in a number of core truths: that AI training requires the reproduction of protected works; that copies of such works are not (only) ephemeral or transitory and are stored in a manner that permits their retrieval for the purpose of producing expressive output that derives from the training data; and that such AI output do, and are likely to, directly compete in the marketplace for expressive works with the works on which the AI was trained, as well as to unfairly displace licensing opportunities that would otherwise exist for authors of the original works.

As we approach how to balance competing interests around AI technologies, we are faced with a litany of arguments that copyright is somehow not fit-for-purpose for this AI. These are the same arguments that were raised with respect to the sound recordings, cable, the photocopier, the internet, and every new technology where one party wanted to make money through uncompensated and unconsented to use of another's creative works.[9]

AI is not as different as the advocates for unfettered, uncompensated reuse pretend.

In my experience in a government relations context, certain arguments against respecting copyright tend to be raised repeatedly. Some of these are made in good faith. For example, where to draw the line on fair use under U.S. law can be the subject of good faith disagreement. Other arguments, such as "we do not make copies," are frankly so factually inaccurate that the speaker risks their credibility. I also place assertions that "all training is per se fair use because it is transformative" in the latter category. U.S. law does not work this way.[10]

So, to level set, here are key points to consider:

---

9. *See* George Thuronyi, *Copyright Law and New Technologies: A Long and Complex* Relationship, LIBR. OF CONG. BLOGS (May 22, 2017), https://blogs.loc.gov/copyright/2017/05/copyright-law-and-new-technologies-a-long-and-complex-relationship/ [https://perma.cc/R47X-GZ3Y] [https://web.archive.org/web/20250318113821/https://blogs.loc.gov/copyright/2017/05/copyright-law-and-new-technologies-a-long-and-complex-relationship/].

10. *See* Jane C. Ginsburg, *Fair Use in the US Redux: Reformed or Still Deformed?*, SING. J. LEGAL STUD. 52, 59–74 (2024) (noting that "[o]ne could espouse a principled position that 'new and important' additions to copyrighted works should not infringe; that the scope of copyright protection should be limited to verbatim, piratical copying. Such a position, however, is not the one Congress chose when it specified exclusive rights over derivative works, and when it directed courts to take into account not only the purpose and character of a defendant's use, but also the amount and substantiality of the use, and the effect of that use upon the potential market for or value of the copyrighted work. Many if not most derivative works 'add something new and important' to the works they copy and adapt; if that were all that was required to render the use 'fair,' then the use 'if it should become widespread, it would adversely affect the potential market for the copyrighted work' by usurping derivative works markets. . . . Looking only at whether the copying of works into training data is a 'transformative' fair use, [the *Andy Warhol Foundation* case] suggests that analysis may depend on whether there is a market for licensing content for training data. Such markets do exist, notably in news media, for high quality, reliable data, and other authors and copyright owners are endeavoring to develop those markets as well. In that event, even if the outputs might not infringe particular inputs, commercial copying (at least) to create training data would be for the same purpose, and might therefore fail a first factor fair use inquiry after [the *Andy Warhol Foundation* case], without a 'compelling justification' for supplanting authors' markets.") (footnotes omitted); *see also* Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169 (2d Cir. 2016) (finding transformative use was still not fair use).

II. LLMS TRAINED ON HIGH-QUALITY MATERIALS, INCLUDING COPYRIGHTED MATERIALS, PRODUCE BETTER OUTCOMES AND CAN FUEL INNOVATION. DISCLOSURE OF TRAINING MATERIALS IS EQUALLY IMPORTANT.

In May 2023, I posted about the importance of disclosure and data quality in AI,[11] borrowing a rubric from assessment, namely high stakes versus low stakes.[12] In that post, I explored the importance of using high quality materials for high stakes applications, and the importance of documenting and disclosing use of such materials:

> When using AI in high stakes decision making, you want to know that your training corpus . . . is pristine and you need to know what is in it. For a pharmaceutical company using AI for decision-making research purposes, the training corpus should be comprised of final Versions of Record (VoR). The researcher needs to know that the corpus excludes unwanted content, such as content sourced from predatory journals and/or "junk" science, for example.
>
> In a low stakes environment there can be a higher tolerance for ambiguity. The same pharmaceutical company researcher may need to simply identify potential experts in a field, which would require a less pristine training corpus; preprints can be included and perhaps even a little "junk" science may be acceptable.[13]

As mentioned in that post, while there can be some value even in "polluted" data, there is a point at which there is too much pollution for most uses. For example, "[b]ias in AI, including racial bias, is well documented."[14] Governments and governmental organizations are moving to regulate AI, focusing on issues such as ethical use and transparency.[15]

---

11. Roy Kaufman, *Swimming in the AI Data Lake: Why Disclosure and Versions of Record Are More Important than Ever*, SCHOLARLY KITCHEN (May 15, 2023), https://scholarlykitchen.sspnet.org/2023/05/15/swimming-in-the-ai-data-lake-why-disclosure-and-versions-of-record-are-more-important-than-ever/ [https://perma.cc/ZC3Q-7B86] [https://web.archive.org/web/20250126035147/https://scholarlykitchen.sspnet.org/2023/05/15/swimming-in-the-ai-data-lake-why-disclosure-and-versions-of-record-are-more-important-than-ever/].

12. *See High-Stakes Test*, GLOSSARY EDUC. REFORM (Aug. 18, 2014), https://www.edglossary.org/high-stakes-testing/ [https://perma.cc/WW2N-YPJF] [https://web.archive.org/web/20250000000000*/https://www.edglossary.org/high-stakes-testing/].

13. Kaufman, *supra* note 11.

14. *Id.* (citing Bernard Marr, *The Problem with Biased AIs (and How to Make AI Better)*, FORBES (Sept. 30, 2022), https://www.forbes.com/sites/bernardmarr/2022/09/30/the-problem-with-biased-ais-and-how-to-make-ai-better/?sh=716485734770 [https://perma.cc/QR6P-WRYF] [https://web.archive.org/web/20250123175140/https://www.forbes.com/sites/bernardmarr/2022/09/30/the-problem-with-biased-ais-and-how-to-make-ai-better/?sh=716485734770]).

15. *Id.* For example, the OECD's ethical AI principles include: "AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art. . . ." *See Transparency and Explainability (Principle 1.3)*, OECD.AI, https://oecd.ai/en/dashboards/ai-principles/P7 [https://perma.cc/FB6S-H5TS] [https://web.archive.org/web/20250123180041/https://oecd.ai/en/dashboards/ai-principles/P7] (last visited Jan. 24, 2025); *see also* Regulation (EU) 2024/1689 of the European Parliament and of the Council of

III. AS MORE CONTENT ONLINE IS AI-GENERATED, ONLINE CONTENT BECOMES LESS RELIABLE

There is a significant body of research indicating that AI, when trained on AI generated content, results in "model collapse."[16] "In the long run, this cycle may pose a threat to A.I. itself. Research has shown that when generative A.I. is trained on a lot of its own output, it can get a lot worse."[17] While it does appear that the effect may be ameliorated by highly controlled efforts in some cases (including closed systems and use of specifically curated and tailored datasets),[18] as a general rule, it argues in favor of ensuring clean supplies of human-generated content are necessary for AI advancement.

---

13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence, and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 2024 O.J. (L. 1689) (imposing transparency requirements on AI developers); Generative AI Copyright Disclosure Act of 2024, H.R. 7913, 118th Cong. § 2(a)(1) (2024) (proposed); N.Y.C., NY ADMIN CODE § 20-871 (2021) (disclosure relating to AI use in hiring).

16. *See* Ilia Shumailov et al., *AI Models Collapse When Trained on Recursively Generated Data*, 631 NATURE 755 (2024); Matyáš Boháček & Hany Farid, *Nepotistically Trained Generative-AI Models Collapse*, ARXIV at 8 (Nov. 20, 2023), https://arxiv.org/pdf/2311.12202 [https://perma.cc/SK2Q-YN5L] [https://web.archive.org/web/20241217072634/https://arxiv.org/pdf/2311.12202] ("We find that at least one popular diffusion-based, text-to-image generative-AI system is surprisingly vulnerable to data poisoning with its own creations. This data poisoning can occur unintentionally . . . Or, it can occur from an adversarial attack . . . ."); Sina Alemohammad et al., *Self-Consuming Generative Models Go MAD*, OPEN REV. (Jan. 16, 2024), https://openreview.net/pdf?id=ShjMHfmPs0 [https://perma.cc/85EY-SQ9S] [https://web.archive.org/web/20250201165329/https://openreview.net/pdf?id=ShjMHfmPs0]; Yunzheng Feng et al., *A Tail of Tails: Model Collapse as a Change of Scaling Laws*. OPEN REV. (Feb. 10, 2024), https://openreview.net/pdf/b07c42e256e6df5c2c52aba4bf28c853110ebb7b.pdf [https://perma.cc/3K28-JV6T] [https://web.archive.org/web/20250219170546/https://openreview.net/pdf/b07c42e256e6df5c2c52aba4bf28c853110ebb7b.pdf]; Quentin Bertrand et al., *On the Stability of Iterative Retraining of Generative Models on Their Own Data*, ARXIV (Apr. 2, 2024), https://arxiv.org/pdf/2310.00429 [https://perma.cc/6AMS-KJJQ] [https://web.archive.org/web/20250126092916/http://arxiv.org/pdf/2310.00429].

17. Aatish Bhatia, *When A.I.'s Output Is a Threat To A.I. Itself*, N.Y. TIMES (Aug. 26, 2024), https://www.nytimes.com/interactive/2024/08/26/upshot/ai-synthetic-data.html [https://perma.cc/5B8K-XD4C] [https://web.archive.org/web/20250123175509/https://www.nytimes.com/interactive/2024/08/26/upshot/ai-synthetic-data.html]; *see also* sources cited *supra* note 16.

18. *See* Yunzhen Feng et al., *Beyond Model Collapse: Scaling Up with Synthesized Data Requires Verification*, ARXIV at 1 (Oct. 25, 2024), https://www.rivista.ai/wp-content/uploads/2024/11/2406.07515v2.pdf [https://perma.cc/D8VG-V3F5] [https://web.archive.org/web/20250219163558/https://www.rivista.ai/wp-content/uploads/2024/11/2406.07515v2.pdf] ("Large Language Models (LLM) are increasingly trained on data generated by other LLM, either because generated text and images become part of the pre-training corpus, or because synthetized data is used as a replacement for expensive human-annotation. This raises concerns about *model collapse*, a drop in model performance when their training sets include generated data. . . . We experiment with two practical tasks—computing matrix eigenvalues with transformers and news summarization with LLMs—which both exhibit model collapse when trained on generated data, and show that verifiers, even imperfect ones, can indeed be harnessed to prevent model collapse and that our proposed proxy measure strongly correlates with performance."); *see also* Lin Long et al., *On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey*, ARXIV (June 14, 2024),

Moreover, as a competitive matter, at the Copyright Clearance Center ("CCC") we have seen a desire by entities training AI systems to license content for training that is not available online. Legality of using publicly posted materials aside, there is no competitive advantage gained from using the same training materials as your competitors.

IV. COPYING TAKES PLACE WHEN CONTENT IS INGESTED INTO LLMS AND AI SYSTEMS. TOKENIZED AND VECTORIZED CONTENT IS STORED AND CAN BE RECALLED.

Copyright law seems complex, but at its core it is quite simple. It is the right to make copies. If copies are made without consent (i.e., a license), it is an infringement unless copying falls under a Berne-permitted copyright exception. So where are copies made?

A recent article co-authored in part by CCC colleagues entitled *The Heart of the Matter* explores the copying that takes place in the process of training AI models:

> LLMs make copies of the documents on which they are trained, and this copying takes various forms, and as a result, with appropriate prompting applications that use the LLMs are able to reproduce original works. The internal representations of the text on which they are trained, in purpose-built vector spaces, are very different in nature from those used in traditional search applications based on indexing because the latter systems consider only the relevance of a given query to the indexed terms of each document, they cannot recreate the indexed documents based on their internal representations—the only way to do this is to actually store a copy of the original text.
>
> It should also be noted that the various forms of copying involve copies that are permanent in nature, such as the initial copies in the training set or the internal representations of the processed text, and transient in nature such as copies made to support the transfer of information between different parts of an AI system or copies related to the output generated during the use of an AI system in what is typically called a "user session."[19]

AI systems "tokenize" words, essentially translating the expressions into numbers. In this process, LLMs make copies of the documents on which they are trained, and this copying takes various forms:

> The process of converting natural language text into a numerical representation involves several steps. The first step is known as "tokenization," which can range from simple separation of words based on whitespace or other separator markings, to more complex techniques like lemmatization and stemming, collectively referred to as "text normalization." Through this process, the natural language text is transformed into a

https://arxiv.org/pdf/2406.15126v1 [https://perma.cc/SG28-ZAZH] [https://web.archive.org/web/20250219160616/https://arxiv.org/pdf/2406.15126v1].

19. Daniel J. Gervais et al., *The Heart of the Matter: Copyright, AI Training, and LLMs*, J. COPYRIGHT SOC'Y (forthcoming) (manuscript at 5) (footnote omitted).

> set of tokens, which are then used to form a "vocabulary"—a list of tokens, each with an associated numerical value.[20]

After being tokenized, text is represented in "embeddings." Word embeddings capture the meaning of a word in the context of the words that surround it as found in the text. In other words, embeddings capture, store, and make use of the author's expression:

> [W]ord embeddings capture, in a dense vector representation, the meaning of a word in the context of the words that surround it, as found in the text that is used during training. So, in essence, word embeddings are a mathematical construct that can efficiently capture the meaning of words based on the various contexts (i.e., word sequences) in which a word can be found. This has been known since the 1950s as the distributional hypothesis. Word embeddings are fundamental blocks during the construction and operation of LLMs . . . .[21]

Numerical representations of the training data that are permanently embedded in LLMs may be considered copies, translations or adaptations of the original works of authorship:

> What the AI models retain post-training are contextual word embeddings that encapsulate the relationships between words in lengthy sequences. The capacity of such systems to reproduce verbatim copies of protected text used as training material, sometimes producing exact or nearly exact copies that are thousands of words long if not more, could be attributed to the fact that these AI systems retain copies, adaptations, or derivative works, stored within the AI systems in specific numerical formats. The fact is that these numerical representations could often be "worked backwards" to recreate a precise and complete version of the original content used as training material.[22]

The fact that some advanced AI systems may include "output filters" to prevent verbatim copies reproduction of works in response to prompts in fact further demonstrates the point: These filters would not be necessary if there were no copies within the systems:

> [I]t should be noted that the fact that some of the more advanced AI systems may be able to install "output filters" that may prevent outputs where large verbatim copies are generated, is of little consequence. As explained above, copies consisting of numerical representations of the training data are made and kept on the AI system regardless of whether the generation of infringing output is regulated at the point of exit.[23]

As summarized by computer scientists Katherine Lee and A. Feder Cooper, with law professor James Grimmelmann, from a legal and technical perspective, "every stage

20. *Id.* at 11 (footnotes omitted).
21. *Id.* at 4.
22. *Id.* at 11 (footnotes omitted).
23. *Id.* at 12–13 (footnotes omitted).

in the generative AI supply chain requires a potentially-infringing reproduction and thus implicates copyright."[24]

Alas, debates persist.

V. TRAINING MATERIALS ADD PERPETUAL BENEFITS TO LLM OUTCOMES.

Once creative output has been used to train an AI system, that system forever benefits.

As former General Counsel of the Copyright Office, Jacqueline Charlesworth, states in her recent article, *Generative AI's Illusory Case for Fair Use*:

> The fair use case for generative AI rests in part on an inaccurate portrayal of the functioning of AI systems. Contrary to the suggestion that the works on which AI systems are trained are set aside after the training process, in fact they have been algorithmically incorporated into and continue to be exploited by the model. AI copying is thus fundamentally different from the copying at issue in the technology-driven fair use precedents relied upon by AI entities. Unlike in these earlier cases—where the copying served functional ends independent of the expressive content of the works—generative AI companies exploit the expressive content of the works they appropriate for its intrinsic value. This exploitation is not confined to the collection of training materials or the training process, but is ongoing and the *sine qua non* of the resulting AI system.[25]

VI. THE USE OF COPYRIGHTED MATERIAL IN THE TRAINING OF LLMS AND IN AI-SYSTEMS SHOULD BE—AND IS—LICENSED: DIRECT AND COLLECTIVE LICENSING

Copying requires licenses. Licensing meets market needs. Unlike the blunt instrument of regulation, it involves parties asking questions such as: "What do you really want?" and "What are you prepared to give/accept?"

Discussing the importance of, and interplay between, direct and collective licensing, *The Heart of the Matter* states:

---

24. Katherine Lee et al., *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*, J. COPYRIGHT SOC'Y (forthcoming 2024) (manuscript at 67); *see also* OpenAI, Comment Letter on the U.S. Patent and Trademark Office's Notice of Inquiry on Intellectual Property Protection for Artificial Intelligence Innovation, at 2, https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf [https://perma.cc/G6R6-2BHM]
[https://web.archive.org/web/20250219214604/https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf] ("Modern AI systems require large amounts of data. For certain tasks, that data is derived from existing publicly accessible "corpora" (singular: "corpus") of data that include copyrighted works. By analyzing large corpora (which necessarily involves first making copies of the data to be analyzed), AI systems can learn . . . .").

25. Jacqueline C. Charlesworth, *Generative AI's Illusory Case for Fair Use*, 27 VAND. J. ENT. & TECH. L. (forthcoming 2025) (manuscript at 4).

> [G]lobal licenses can harmonize how copyright owners and users agree to use copyrighted works, significantly benefiting innovation and progress by setting the stage for consistent and responsible copyright uses that could lead to untold scientific and cultural advancements. Licenses could put an end to much of the uncertainty and to both pending and potential future litigation, putting acceptable boundaries on what can and cannot be done with copyrighted material when training LLMs.
>
> Various licensing models could play a crucial role in this progress. Direct licensing—agreements between a copyright owner and user—is incredibly important because it allows the parties to be flexible in defining terms like payment, timing, and addressing specific, bespoke use cases. Voluntary collective licensing is also likely to play a critical role in solving the licensing puzzle, enabling users to obtain a single license that can cover thousands (or more) of copyrighted works without having to negotiate with each copyright owner individually. This approach is highly beneficial for both copyright owners and users, as it provides an efficient mechanism to grant and obtain permission for using copyrighted works.
>
> Voluntary collective licensing is uniquely equipped to handle some of the more complex issues, when there are large numbers of works and potential users searching for an efficient mechanism to provide and obtain permission for using copyrighted works. One example of how this might be helpful in the AI context is a company engaged in heavy research and development activities that may want to make additional internal uses of a large number of textual works that they have acquired lawfully. The company may not have the bandwidth to engage in additional negotiations, while the publishers of the various scholarly journals would similarly be interested in licensing but would prefer to rely on a more streamlined approach. Importantly, voluntary collective licenses complement direct licenses, providing a framework where copyright owners and users can rely on collective licenses for many typical use cases and direct licenses for unique or individualized situations.
>
> In the case of AI, we believe that both direct and collective licenses can be valuable to reduce uncertainty and establish a viable ecosystem going forward. Some uses, such as certain training activities or general categories of outputs that need access to diffuse copyrighted materials, may be good candidates for collective licensing. Conversely, specific high-value or individual uses based on more defined sets of copyrighted materials could be better suited for direct licensing. Regardless of the approach, licensing provides both parties with compliant access to high-quality works, leading to innovative uses.[26]

VII. CCC AND AI LICENSING

CCC's history, from its founding in 1978,[27] has reduced licensing friction in response to new technologies.

---

26. Gervais et al., *supra* note 19, at 27–28 (footnotes omitted).

27. COPYRIGHT CLEARANCE CTR., CREATING SOLUTIONS TOGETHER; LESSONS TO INFORM THE FUTURE OF COLLECTIVE LICENSING 32 (2020), https://www.copyright.com/wp-content/uploads/2021/01/CCC_CreatingSolutionsTogether_Ebook_2020.pdf [https://perma.cc/RR7X-T262] [[https://web.archive.org/web/20250312172318/https://www.copyright.com/wp-content/uploads/2021/01/CCC_CreatingSolutionsTogether_Ebook_2020.pdf].

There was a time when photocopying was deemed disruptive. Congress considered whether changes were required to make photocopies "fair use" under U.S. law.[28] Congress no doubt recognized that photocopying, like AI, is a tool, not a "use" capable of being declared categorically "fair" or "infringing." Thus, in *Basic Books v. Kinko's*, the court did not decide liability on whether a photocopier was used, but on whether the photocopies in question were being used for commercial or non-commercial purposes.[29] The court there looked not just at the ultimate use of the copies—which was for students in the classroom—but also at the defendant's use, which was distinct from that of the students.[30] Kinko's use was deemed commercial and infringing.[31]

CCC offers market-based, global non-exclusive voluntary licenses to support AI in commercial research, schools, and education technology sectors. These licenses were built with rightsholders and users based on agreed understandings of needs and market conditions.

For example, in July 2024 we announced the inclusion of AI rights into our Annual Copyright License ("ACL") for internal use by corporations.[32] The inclusion of AI re-use rights, based largely on demand from our corporate customers, was the first-ever collective licensing solution specifically designed specifically for AI. As of today, these rights cover hundreds of thousands of users globally.

Prior to the addition of specifically denominated "AI rights" in the ACL, we offered other licenses to support specific AI-relevant use cases,[33] including licenses designed for text and data mining,[34] and to allow the creation of machine generated materials for use in classrooms[35] Like the ACL, these rights were added at the request of users who approached CCC with a use case which needed a license. This is how licensing works:

---

28. *Id.* at 25.

29. *See* Basic Books, Inc. v. Kinko's Graphics Corp., 758 F. Supp. 1522, 1530 (S.D.N.Y 1991).

30. *Id.* at 1532.

31. *Id.* at 1531.

32. *See CCC Pioneers Collective Licensing Solution for Content Usage in Internal AI* Systems, COPYRIGHT CLEARANCE CTR. (July 16, 2024), https://www.copyright.com/media-press-releases/ccc-pioneers-collective-licensing-solution-for-content-usage-in-internal-ai-systems/ [https://perma.cc/UNQ2-HLAA].

33. *See, e.g.*, *Unlock the Value of Scientific Literature Using Machine-Readable Articles*, COPYRIGHT CLEARANCE CTR., https://www.copyright.com/solutions-rightfind-xml/ [https://perma.cc/MK3R-W32L] [https://web.archive.org/web/20250318123427/https://www.copyright.com/web/20250318123427/https://www.copyright.com/solutions-rightfind-xml/] (last visited Mar. 18, 2025); *Easily Search for, Discover, and Incorporate High-Quality, Copyrighted Content into Curriculum and Instruction*, COPYRIGHT CLEARANCE CTR., https://www.copyright.com/solutions-annual-copyright-license-for-curriculum-instruction/ [https://perma.cc/7B3J-2N3M] [https://web.archive.org/web/20250318123841/https://www.copyright.com/solutions-rightfind-curriculum/] (last visited Mar. 18, 2025). These licenses do not use the phrase "AI," partly because they predate its common usage and partly because they support more specific use cases.

34. *See Unlock the Value of Scientific Literature Using Machine-Readable Articles*, COPYRIGHT CLEARANCE CTR., https://www.copyright.com/solutions-rightfind-xml/ [https://perma.cc/B9PK-73ST] [https://web.archive.org/web/20250219222119/https://www.copyright.com/solutions-rightfind-xml/] (last visited Jan. 24, 2025).

35. *See Annual Copyright License for Curriculum & Instruction*, COPYRIGHT CLEARANCE CTR., https://www.copyright.com/solutions-annual-copyright-license-for-curriculum-instruction/ [https://perma.cc/62VX-N4NH] (last visited Jan. 24, 2025).

A willing buyer seeks a willing seller. Many AI developers skipped this step, preemptively declaring "licensing is not possible."[36]" That is not correct.

In addition to collective licenses, rights owners are licensing content for use in AI systems transactionally on a global basis. New transactional AI deals are regularly announced,[37] and many more deals are closed and kept confidential. CCC is involved in this market too.

VIII. OPEN LICENSING AND AI

In addition to traditional transactions and collective licensing models, AI use is also licensed under so called "open models" such as those offered by Creative Commons ("CC"),[38] and GNU licenses.[39] For example, science publishing uses a variety of "open access" business models.[40] These models are based, in part, on open licensing:

> We encourage the use of CC licenses, not only because they are very well established legal tools, but because they have the benefits of simplicity, machine-readability and interoperability. Importantly, many elements of internet infrastructure 'understand'

36. *See, e.g.*, Microsoft, Reply Comment Letter on the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, at 9 (Oct. 30, 2023), https://www.regulations.gov/comment/COLC-2023-0006-8750 [https://perma.cc/C3QW-RDFE] [https://web.archive.org/web/20250313201720/https://www.regulations.gov/comment/COLC-2023-0006-8750] ("Any requirement to obtain consent for accessible works to be used for training would chill AI innovation. It is not feasible to achieve the scale of data necessary to develop responsible AI models even when the identity of a work and its owner is known.")

37. *See Generative AI Licensing Agreement Tracker*, ITHAKA S+R, https://sr.ithaka.org/our-work/generative-ai-licensing-agreement-tracker/ [https://perma.cc/T3DH-76A3] [https://web.archive.org/web/20250124164347/https://sr.ithaka.org/our-work/generative-ai-licensing-agreement-tracker/] (last visited Jan. 24, 2025); Paul Sweeting, *Generative AI & Licensing, A Special Report*, VARIETY (Oct. 1, 2024), https://variety.com/vip-special-reports/generative-ai-content-licensing-special-report-1236157051/ [https://perma.cc/BX5D-4ABV] ("Alongside the rapid rise of generative AI, a new licensing market has begun to emerge for AI training data. Since mid-2023, AI companies have been pursuing licensing deals with media rights holders to secure access to their content and use it as high-quality data to train powerful AI models of any modality, notably including text, image, music and video. To date, more than two dozen content owner deals with AI developers have been publicly confirmed, according to VIP+ research. A diverse range of publisher types are now engaged in licensing, with dealmaking rampant among news publishers, stock image companies and platforms such as Reddit and Stack Overflow.").

38. *See Homepage*, CREATIVE COMMONS, https://creativecommons.org/ [https://perma.cc/5U4E-RV9X] [https://web.archive.org/web/20250124120047/https://creativecommons.org/] (last visited Jan. 24, 2025).

39. *See Licenses*, GNU OPERATING SYS., https://www.gnu.org/licenses/licenses.en.html [https://perma.cc/FF4T-D255] [https://web.archive.org/web/20250124171254/https://www.gnu.org/licenses/licenses.en.html] (last visited Jan. 24, 2025).

40. *See 2023 OA Progress Report*, INT'L ASS'N SCI., TECH. & MED. PUBLISHERS (2023), https://s3.eu-west-2.amazonaws.com/stm.offloadmedia/wp-content/uploads/2024/08/10032807/2023OAProgressReportFINAL-2-1.pdf [https://perma.cc/2S3A-H8KZ] [https://web.archive.org/web/20250124172023/https://s3.eu-west-2.amazonaws.com/stm.offloadmedia/wp-content/uploads/2024/08/10032807/2023OAProgressReportFINAL-2-1.pdf].

> CC licensing, and can display and filter content appropriately, based on this machine-readable license information . . . .[41]

Use of open licenses in AI must, of course, comply with the relevant license terms.[42]

IX. CONCLUSION

In the esoteric field of copyright, people tend to gravitate to rhetorical clashes between past and present, between man and machine. The reality is infinitely more boring and quotidian. Reality reduces to the straightforward issue of licensing, and here, we do not need to reinvent the wheel. We should reject claims of unprecedented complexity and remain focused on reality. Copyright—while it may provide authors with the right to prevent uses—is largely about the facilitation of licensing. Rather than imagining ourselves caught in an existential battle over the future, we should review how the past informs where we go from here. Bluntly, licensing has solved seemingly complex reuse challenges in the past, enables AI, and is fit-for-purpose for the future.

Let us eschew drama and reject specious advocacy that somehow "this time is truly different" and we need to support "innovators" by suspending consent of creators -- who are innovators themselves. Such claims tend to be made with every technological development, and they universally oversimplify narratives in favor of one stakeholder group—a particularly wealthy stakeholder group which has chosen to free ride on others. We can, and have always, made technological progress while respecting the rights of creators. We do not need to reinvent law or our approach to licensing. The challenge is not unprecedented, and we must avoid taking unprecedented and dangerous steps. Every day, responsible AI companies and willing rightsholders are demonstrating that we can build AI models through voluntary licensing—licensing that can be adapted as we learn more about the implications of generative AI on other uses of copyright works. So let us follow the path of reason. We can sustain technological innovation while protecting authors and other creators. The answer, as always, comes back to one thing…licensing.

To quote the U.S. Supreme Court in *Andy Warhol Foundation v. Goldsmith*:

> It will not impoverish our world to require [defendant] to pay [artist] 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. Nor will the Court's decision, which is consistent with longstanding principles of fair use, snuff out the

41. *FAQs*, OPEN ACCESS SCHOLARLY PUBL'G ASS'N, https://www.oaspa.org/about/faqs/ [https://perma.cc/6QJ9-LGZJ] [https://web.archive.org/web/20250124172551/https://www.oaspa.org/about/faqs/] (last visited Jan. 24, 2025).

42. *See* Roy Kaufman, *GitHub Is Sued, and We May Learn Something About Creative Commons Licensing*, SCHOLARLY KITCHEN (Jan. 5, 2023), https://scholarlykitchen.sspnet.org/2023/01/05/github-is-sued-and-we-may-learn-something-about-creative-commons-licensing/ [https://perma.cc/JRX5-4CZ9] [https://web.archive.org/web/20250124173016/https://scholarlykitchen.sspnet.org/2023/01/05/github-is-sued-and-we-may-learn-something-about-creative-commons-licensing/].

> light of Western civilization, returning us to the Dark Ages of a world without Titian, Shakespeare, or Richard Rodgers.[43]

Copyright exists to reward creativity and encourage the creation of new materials. Or, as stated in Article I, Section 8, Clause 8 of the U.S. Constitution, "To promote the Progress of Science and useful Arts."[44] It has done so for centuries. Let's support AI by supporting copyright, not by destroying it.

43. Andy Warhol Found. Visual Arts v. Goldsmith, 549 U.S. 508, 549 (2023). For a discussion of the implications of *Warhol* for AI-related copyright cases, see Roy Kaufman, *The Supreme Court Case of* Andy Warhol Foundation v. Goldsmith*: What, if Anything, Does It Mean To Artificial Intelligence?*, SCHOLARLY KITCHEN (June 6, 2023), https://scholarlykitchen.sspnet.org/2023/06/06/the-supreme-court-case-of-andy-warhol-foundation-v-goldsmith-what-if-anything-does-it-mean-to-artificial-intelligence/ [https://perma.cc/H7Q2-RTGS] [https://web.archive.org/web/20250124175244/https://scholarlykitchen.sspnet.org/2023/06/06/the-supreme-court-case-of-andy-warhol-foundation-v-goldsmith-what-if-anything-does-it-mean-to-artificial-intelligence/].

44. U.S. CONST. art. I, § 8, cl. 8.