Consent and Compensation:
Resolving Generative AI's Copyright Crisis

Frank Pasquale
Haochen Sun

Generative artificial intelligence (AI) has the potential to augment and democratize creativity. However, it is undermining the knowledge ecosystem that now sustains it. Generative AI may unfairly compete with creatives, displacing them in the market. Most AI firms are not compensating creative workers for composing the songs, drawing the images, and writing both the fiction and non-fiction books that their models need in order to function. AI thus threatens not only to undermine the livelihoods of authors, artists, and other creatives, but also to destabilize the very knowledge ecosystem it relies on.

Alarmed by these developments, many copyright owners have objected to the use of their works by AI providers. To recognize and empower their demands to stop non-consensual use of their works, we propose a streamlined opt-out mechanism that would require AI providers to remove objectors' works from their databases once copyright infringement has been documented. Those who do not object still deserve compensation for the use of their work by AI providers. We thus also propose a levy on AI providers, to be distributed to the copyright owners whose work they use without a license. This scheme is designed to ensure creatives receive a fair share of the economic bounty arising out of their contributions to AI. Together these

mechanisms of consent and compensation would result in a new grand bargain between copyright owners and AI firms, designed to ensure both thrive in the long-term.

Cite as: Frank Pasquale and Haochen Sun, *Consent and Compensation: Resolving Generative AI's Copyright Crisis*, 110 U. VA. L. REV. ONLINE (forthcoming, 2024).

Consent and Compensation:
Resolving Generative AI's Copyright Crisis

Frank Pasquale*

Haochen Sun**

I. INTRODUCTION

From the printing press to the Internet, technological advance has profoundly changed the way authors create, disseminate, and monetize their works.[1] Widespread access to the Internet has caused book, music, and film creators great economic setbacks via piracy, but has also created new opportunities, particularly for "long tail" creators shunned by dominant recording companies and broadcasters.[2] Despite the upheaval, human authors have remained indispensable in the creation of works, as pirates do not create original content.

The rise of generative artificial intelligence (AI), however, represents an inflection point.[3] AI can plagiarize at a far faster rate than

* Professor of Law, Cornell Law School and Cornell Tech.

* Professor of Law, University of Hong Kong Faculty of Law. We thank Shyam Balganesh, Anupam Chander, William Fisher, James Grimmelmann, Jacob Noti-Victor, Ben Sobel, Scott Veitch, and Christopher Yoo for valuable comments and conversations. We are grateful to participants at the Hong Kong University conference *Reframing Intellectual Property Law in the Age of Artificial Intelligence* for their comments in response to a presentation of this project. We also thank Michelle Brodsky, Alex Cho, Jae Shin, and Upasana Singh for excellent research assistance.

[1] See generally Adrian Johns, Piracy: The Intellectual Property Wars from Gutenberg to Gates (2009) (discussing the history of copyright piracy).

[2] See Chris Anderson, The Long Tail, Wired (Oct. 1, 2004, 12:00 PM), https://www.wired.com/2004/10/tail [https://perma.cc/P9QQ-MPTG].

[3] Generative AI's power to create exact replicas of existing works, and to imitate many characteristic elements of existing work, has provoked a wave of lawsuits over the past two years. However, copyright controversies over the training of AI antedate the rise of generative AI. To mark the relevance of that past work, and the continuity of the problems likely to be raised by AI when the next generation of AI arises, we refer to "AI" throughout the article, rather than the more cumbersome "generative AI" or "GenAI."

human copyists.[4] These capacities are menacing both fiction and non-fiction book authors and journalists.[5] AI can also create new works that closely resemble the style and content of existing ones. When prompted skillfully, large language models (LLMs) aid in the rapid creation of a high volume of works. The bottom line is an "existential crisis" for many creatives, threatening to drive the marginal value of their labor below subsistence levels as cheap AI content displaces human works.[6]

Given the enthusiasm for AI evident among so many owners of dominant content distribution platforms, such a displacement may already be underway.[7] To create and improve their AI models, large technology firms have undermined authors' proprietary control over their works by using these works as training data, without consent and often through opaque processes.[8] At the same time, AI systems like ChatGPT and MidJourney can rapidly generate a wide variety of content, potentially outperforming humans in the marketplace of ideas—particularly when so many of this marketplace's main

[4] Kate Knibbs, Scammy AI-Generated Book Rewrites Are Flooding Amazon, Wired (Jan. 10, 2024, 7:00 AM), https://www.wired.com/story/scammy-ai-generated-books-flooding-amazon/ [https://perma.cc/4R7G-LXFU].

[5] Our focus in this essay is on corporations developing, marketing, and selling AI services. The legislative approaches developed in this essay may, in a calibrated fashion, adjust duties of AI providers to reflect their size, for-profit or non-profit status, and other factors.

[6] See Michael Cavna, Artists Are Alarmed by AI—and They're Fighting Back, Wash. Post (Feb. 14, 2023, 6:00 AM), https://www.washingtonpost.com/comics/2023/02/14/ai-in-illustration/ [https://perma.cc/4RFW-5FX3**]** (describing "an existential threat to the livelihood of artists"). Throughout this essay, we will refer to artists, writers, journalists, and other creators of expressive works as "creatives" or "copyright owners." We realize these terms may be too capacious: some expressive work only takes a minimal amount of creativity, and many creatives have transferred their copyrights to others in exchange for compensation. Nevertheless, copyright is premised on some minimal level of creativity, and the future compensation of creatives who plan to alienate their copyrights is at least in part premised on the value of those copyrights to those seeking them. So, the terms capture enough of social and economic reality to be useful here.

[7] Edward Zitron, Are We Watching The Internet Die?, Where's Your Ed At? (Mar. 11, 2024), https://www.wheresyoured.at/are-we-watching-the-internet-die/ [] (recognizing that because "platforms were built to reward scale and volume far more often than quality," creatives using AI enjoy important advantages over those who do not.)

[8] See infra Part II.B.

organizers such as Alphabet (formerly Google), X (formerly Twitter), and Meta (formerly Facebook) are themselves developing AI.[9]

To compound these challenges, leading firms in the AI space are unlikely to offer compensation for the vital contribution of copyrighted works to their systems. In 2023, this state of affairs helped lead to an unprecedented 148-day strike by Hollywood screenwriters.[10] Book authors are also alarmed. Over 15,000 writers, including prominent novelists such as Dan Brown, Suzanne Collins, and Margaret Atwood, have endorsed an open letter demanding fair compensation, credit, and author consent.[11] At least one former executive in an AI firm has resigned his position, considering the unlicensed use of music as training data both ethically and legally untenable.[12] This struggle has resulted in numerous courtroom battles over copyright infringement, too.[13] AI firms claim that they are

[9] See Thomas H. Davenport & Nitin Mittal, How Generative AI Is Changing Creative Work, Harv. Bus. Rev. (Nov. 14, 2022), https://hbr.org/2022/11/how-generative-ai-is-changing-creative-work [https://perma.cc/SK98-ZE5T].

[10] Ben Schwartz, AI and the Hollywood Writers' Strike, The Nation (May 8, 2023), https://www.thenation.com/article/economy/ai-and-the-hollywood-writers-strike [https://perma.cc/8TJR-ZBUC]; Jennifer Maas, The Writers Strike Is Over: WGA Votes to Lift Strike Order After 148 Days, Variety (Sept. 26, 2023, 5:07 PM) https://variety.com/2023/tv/news/writers-strike-over-wga-votes-end-work-stoppage-1235735512/ [https://perma.cc/F5P7-QEWF].

[11] Open Letter to Generative AI Leaders, Action Network, https://actionnetwork.org/petitions/authors-guild-open-letter-to-generative-ai-leaders [https://perma.cc/8D5W-WGFL] (last visited Mar. 3, 2024).

[12] Kate Knibbs, This Tech Exec Quit His Job To Fight Generative AI's Original Sin, Wired (Jan. 17, 2024, 4:44 PM), https://www.wired.com/story/ai-executive-ed-newton-rex-turns-crusader-stand-up-for-artists [https://perma.cc/97NE-H4Y7].

[13] Complaint at 2–3, Basbanes v. Microsoft Corp., No. 24-cv-00084 (S.D.N.Y. Jan. 5, 2024); Complaint at 2–4, N.Y. Times Co. v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. Dec. 27, 2023); Generative AI-Intellectual Property Cases and Policy Tracker, Mishcon de Reya LLP, https://www.mishcon.com/generative-ai-intellectual-property-cases-and-policy-tracker [https://perma.cc/7RHU-3PG2] (last visited Mar. 3, 2024).

protected by the fair use defense,[14] but application of the doctrine is notoriously uncertain, particularly with respect to new technologies.[15]

This litigation may drag on for years, slowing the development of AI while denying or delaying fair compensation to creatives. The situation strikes many policymakers as deeply unfair and undesirable. As the Communications and Digital Committee of the United Kingdom's House of Lords has concluded, "We do not believe it is fair for tech firms to use rightsholder data for commercial purposes without permission or compensation, and to gain vast financial rewards in the process."[16] A legislative solution is desirable, and there is a venerable tradition of actual and proposed solutions to the copyright problems created by new technological uses of works.[17]

To guide policymakers, this essay outlines a promising framework for a legislative solution, premised on coupling mechanisms of control (via opt-out rights) and compensation (via a levy to be imposed on AI providers by a central authority, and then distributed to owners of works used by those AI providers without a license). These mechanisms could first be imposed on the largest AI providers, and then expanded as appropriate once standardized. Part II explains the urgency of this proposal by demonstrating that free expropriation of copyrighted works by AI providers not only devalues human creativity but also threatens to undermine AI itself by

[14] Mark A. Lemley & Bryan Casey, Fair Learning, 99 Tex. L. Rev. 743, 748 (2021) (arguing that "an [machine learning] system's use of the data often *is* transformative as that term has come to be understood in copyright law, because even though it doesn't change the underlying work, it changes the purpose for which the work is used") (emphasis in original).

[15] Katherine Lee, A. Feder Cooper & James Grimmelmann, Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain, 71 J. Copyright Soc'y (forthcoming 2024) (manuscript at 105), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4523551 [https://perma.cc/Z3C7-PJWJ] ("[F]air use is famously case-specific, so no *ex ante* analysis can anticipate all of the relevant issues.").

[16] Commc'ns & Digit. Comm., Large Language Models and Generative AI, 2023-24, HL 54, ¶ 245 (UK).

[17] See William W. Fisher III, Promises to Keep: Technology, Law, and the Future of Entertainment 1–22 (2004).

eliminating critical incentives for the ongoing creation of works necessary for its advance. Part III outlines an opt-out mechanism, permitting creatives to forbid non-consensual use of their works for training AI models after documenting copyright infringement. Part IV addresses the proper level of levies necessary to compensate those who do not choose to opt out or license their works to AI providers. Part V anticipates and responds to objections to our proposal, while Part VI concludes by reflecting on its broader policy implications.

II. AI'S COPYRIGHT CRISIS

Myriad texts and images inform the models powering apps like ChatGPT and DALL-E. AI is ultimately parasitic on training data. Many parasitic relationships exist in stable equilibria throughout the natural and economic world; however, sometimes a parasite can overwhelm its host. This is a pressing danger in the new digital knowledge ecosystem, as explained in Sections A and B below. Section C then explores how AI may harm the quality of the training data it needs if it sufficiently undercuts creatives with cheap and prolific outputs unmoored from direct human observation and experience.

A. Copyright, Consent, and the Knowledge Ecosystem

Copyright law plays an essential role in the knowledge ecosystem. It encourages authors to create works by granting them exclusive rights.[18] These rights entitle authors to prevent others from reproducing, distributing, or publicly performing their works without

[18] The Intellectual Property Clause of the U.S. Constitution grants Congress the enumerated power "[t]o promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." U.S. Const. art. I, § 8, cl. 8. *See also* Mazer v. Stein, 347 U.S. 201, 219 (1954) (concluding that copyright law incentivizes authors by granting exclusive rights, in order "to afford greater encouragement to the production of literary [or artistic] works of lasting benefit to the world").

permission.[19] By granting exclusive rights, the copyright system incentivizes creatives to publish their works. It also reduces piracy and unwanted derivative works.[20] The exclusivity of copyright forms the legal basis for authors' proprietary control over their creations.[21] It allows them to protect their works from unauthorized access and use and to grant permissions for access and use, often in exchange for financial rewards such as royalties.[22] Copyright law also incentivizes many intermediaries to disseminate creators' works.[23] It therefore awards publishers, performers, and broadcasting organizations with a range of related rights for their contributions to disseminating works to the public. Hence, authors' control of works, and compensation for them, are central to the knowledge ecosystem. Copyright law empowers authors to not only give consent for the use and access of their works but also to receive compensation associated with such permissions, subject to limitations such as fair use.[24]

The opacity and scale of AI systems is disrupting the knowledge ecosystem by significantly eroding authors' proprietary control of their works, well beyond extant digital practices that have already undermined many authors' well-being. Whereas prior scraping at scale tended to be focused on the non-expressive aspects of works (such as facts), AI is focused by many prompts on their expressive dimensions. Search engines have historically provided links which lead

[19] Jeanne C. Fromer & Christopher Jon Sprigman, Copyright Law: Cases and Materials 213 (3d ed. 2021) (listing exclusive rights); Neil Weinstock Netanel, Copyright and a Democratic Civil Society, 106 Yale L.J. 283, 285 (1996) ("To encourage authors to create and disseminate original expression, copyright law accords them a bundle of proprietary rights in their works.").
[20] Mark A. Lemley, Property, Intellectual Property, and Free Riding, 83 Tex. L. Rev. 1031, 1059 (2005).
[21] Robert P. Merges, Justifying Intellectual Property 5 (2011) (highlighting "individual control" over intangible assets as a core principle of IP law).
[22] Shyamkrishna Balganesh, Foreseeability and Copyright Incentives, 122 Harv. L. Rev. 1569, 1578 (2009).
[23] Id. at 1622–23.
[24] Merges, supra note 21, at 197 (arguing that the main contribution of IP protection is "augmenting income" for creatives).

users to works themselves. In contrast, AI tends to provide substitutes for such works, while failing to provide citations to the works in the dataset most similar to the texts, images, and videos it presents as a computed synthesis.

Training AI requires use of large volumes of copyrighted works without obtaining authorization from their authors, bypassing human authors' control in two ways. First, in pursuit of high-quality datasets, AI developers have deliberately targeted copyrighted materials at scale. Many AI providers have ignored the "robots.txt" convention that, for many years, permitted website owners to opt out of many forms of large-scale web-scraping with minimal effort.[25] Books3, a dataset comprising nearly 200,000 copyrighted e-books, has been employed to train AI systems operated by companies like Meta and Bloomberg.[26] This diverse dataset is valuable for training purposes, as it includes books from various genres, ranging from obscure erotic fiction and poetry to acclaimed novels by well-known authors (including Stephen King and Margaret Atwood).[27] Given the secrecy of AI firms' operations, it is unclear whether they made any effort to obtain permission from these authors. However, the thousands of authors who signed on to a letter complaining about this use of their work is good circumstantial evidence that permission was not sought.[28]

Denounced as "the biggest act of copyright theft in history" and "unbelievably disrespectful," the use of Books3 for data training has

[25] David Pierce, The Text File that Runs the Internet, The Verge (Feb. 14, 2024, 9:00 AM), https://www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders [https://perma.cc/K4MH-U56X] ("For decades, robots.txt governed the behavior of web crawlers. But as unscrupulous AI companies seek out more and more data, the basic social contract of the web is falling apart.").

[26] Alex Reisner, What I Found in a Database Meta Uses to Train Generative AI, The Atlantic (Sept. 25, 2023), https://www.theatlantic.com/technology/archive/2023/09/books3-ai-training-meta-copyright-infringement-lawsuit/675411 [https://perma.cc/58V5-2NVP].

[27] Id.

[28] See Open Letter to Generative AI Leaders, supra note 11.

provoked anger, frustration, and fear among authors.[29] One claimed their "soul had been strip mined" and they felt "powerless to stop it."[30] Another described being "completely gutted and whipsawed."[31] Some leading AI researchers have also objected; for example, Australian computer scientist Toby Walsh has repeatedly criticized the use of Books3.[32] Moreover, this exploitation extends beyond literary works: there are also numerous images exemplifying AI's "visual plagiarism problem."[33]

Second, some AI providers have themselves scraped a huge trove of works from the Internet, while others have utilized intermediaries to gain access to works. Consider, for instance, the landscape of text-to-image generation: "While Stable Diffusion and its variants have been trained on open-sourced datasets. . . little is known about the datasets that are used to train models such as OpenAI's Dall-E, Google's Parti, and Imagen."[34] One of these open-sourced datasets, the Large-Scale Artificial Intelligence Open Network (LAION), has provided access to billions of training images as of October 2022,

[29] Kelly Burke, 'Biggest Act of Copyright Theft in History': Thousands of Australian Books Allegedly Used to Train AI Model, The Guardian (Sept. 28, 2023, 11:00 AM), https://www.theguardian.com/australia-news/2023/sep/28/australian-books-training-ai-books3-stolen-pirated [https://perma.cc/L5YV-4W73]; Valerie Ouellet, Sylvène Gilchrist & Shaki Sutharsan, CBC News Analysis Finds Thousands of Canadian Authors, Books in Controversial Dataset Used to Train AI, CBC News (Dec. 7, 2023, 4:00 AM), https://www.cbc.ca/news/canada/canadian-authors-books3-ai-dataset-1.7050243 [https://perma.cc/J8RN-6RTB].

[30] Burke, supra note 29.

[31] Leah Asmelash, These Books Are Being Used to Train AI. No One Told the Authors, CNN (Oct. 8, 2023, 8:00 AM), https://edition.cnn.com/2023/10/08/style/ai-books3-authors-nora-roberts-cec/index.html [https://perma.cc/L3J9-7H5R].

[32] See, e.g., Toby Walsh (@TobyWalsh), X (Jan. 20, 2024, 10:01 PM), https://twitter.com/TobyWalsh/status/1748903611311313275 [https://perma.cc/MA6X-BVVZ].

[33] Gary Marcus & Reid Southen, Generative AI Has a Visual Plagiarism Problem: Experiments with Midjourney and DALL-E 3 Show a Copyright Minefield, IEEE Spectrum (Jan. 6, 2024), https://spectrum.ieee.org/midjourney-copyright [https://perma.cc/F4WW-7XNS].

[34] Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti & Alexandra Sasha Luccioni, Into the LAION's Den: Investigating Hate in Multimodal Datasets, 2 (Nov. 6, 2023) (unpublished manuscript), https://arxiv.org/pdf/2311.03449.pdf [https://perma.cc/4HHX-4KU7].

making it the largest image dataset for training machine-learning models.[35] LAION's processes of image aggregation do not appear to include seeking permission from copyright owners.[36] Instead, the network generates image-text pairs by first utilizing Common Crawl's metadata files, extracting URLs of images with captions, and then downloading the raw images from the parsed URLs.[37]

While LAION boasts non-profit status, it is supported by and in turn supports several for-profit firms which use its datasets for commercial purposes.[38] OpenAI employs GPTbot, a powerful web crawler, to scrape and collect virtually any online content for AI model training. Consequently, the upcoming GPT-5 model will likely be trained on copyrighted content gathered by this bot without permission from rights holders.[39]

In response, thousands of artists, writers, designers, and photographers have posted "Do Not AI" signs on their social media

[35] Romain Beaumont et al., LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models 1 (Mar. 31, 2022) (unpublished manuscript), https://arxiv.org/pdf/2210.08402.pdf [https://perma.cc/U3D6-NB2U].

[36] Admittedly, in some cases LAION is merely aggregating images from extant aggregators which themselves paid scant attention to copyright. However, it is much easier for creatives with valid copyright claims to utilize notice and takedown measures with respect to those aggregators, than it is to request LAION to keep links to works out of its dataset. Chloe Xiang, A Photographer Tried to Get His Photos Removed from an AI Dataset. He Got an Invoice Instead., Motherboard (Apr. 28, 2023, 9:00 AM), https://www.vice.com/en/article/pkapb7/a-photographer-tried-to-get-his-photos-removed-from-an-ai-dataset-he-got-an-invoice-instead [https://perma.cc/ZXV2-PXMP] ("A German stock photographer who asked to get his images removed from a dataset used to train AI image generators was not only met with a refusal from the dataset owner but also an invoice for $979 for filing an unjustified copyright claim.").

[37] Romain Beaumont et. al, LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b [https://perma.cc/FDF4-BWLY].

[38] Kyle Chayka, Is A.I. Art Stealing from Artists?, The New Yorker (Feb. 10, 2023), https://www.newyorker.com/culture/infinite-scroll/is-ai-art-stealing-from-artists [https://perma.cc/A9NA-5BU5].

[39] Alistair Barr, OpenAI Just Admitted It Has a Bot That Crawls the Web to Collect AI Training Data. If You Don't Block GPTbot, That's Self-Sabotage., Bus. Insider (Aug. 8, 2023, 5:10 PM), https://www.businessinsider.com/openai-gptbot-web-crawler-content-creators-ai-bots-2023-8 [https://perma.cc/N8TG-KZXX].

accounts, protesting the use of their works for AI model training.[40] However, due to the black-box nature of AI systems, it can be difficult for creatives to determine if their works have been used for such purposes.[41]

Despite these grave threats to the livelihoods of creatives, AI providers often deny the need to obtain consent to use their works. Indeed, it is difficult to predict how courts will rule on many copyright owners' infringement claims against AI providers. Consider, first, the question of infringement itself. At least at the production phase, AI providers will likely claim that they are a mere tool of their users, who should be responsible for infringement if they prompted the AI to create an infringing work. However, even if users are held liable for direct infringement, AI providers could still be vicariously or contributorily liable for infringement they enable.

Menaced by such secondary liability claims, AI providers will tend to portray their service as having substantial non-infringing uses, citing favorable precedents regarding the video cassette recorder (VCR). The Supreme Court upheld the legality of the VCR device, because it was capable of "substantial noninfringing uses," including fair use of copyrighted work by VCR owners who time-shifted their viewing of broadcast television programs by taping them and watching them later.[42] Nevertheless, there is a key difference between AI as a

[40] Gayanga Dissanayaka, AI: A Threat or a Tool for Creative Fields?, Daily Mirror (June 1, 2023, 12:10 AM), https://www.dailymirror.lk/news-features/AI:-A-Threat-or-a-Tool-for-Creative-Fields-/131-260217 [https://perma.cc/ZF9R-4CBS].

[41] The Daily, The Writers' Revolt Against A.I. Companies, N.Y. Times, at 7:13 (July 19, 2023), https://www.nytimes.com/2023/07/18/podcasts/the-daily/ai-scraping.html [https://perma.cc/E327-Z7TH]..

[42] Sony Corp. of Amer. v. Universal City Studios, Inc., 464 U.S. 417, 454–56 (1984); id. at 450 n.33 ("[T]he time-shifter no more steals the program by watching it once than does the live viewer, and the live viewer is no more likely to buy prerecorded videotapes than is the time-shifter. Indeed, no live viewer would buy a prerecorded videotape if he did not have access to a VTR."). As one of us has argued in past work,

> The majority offer[ed] no empirical evidence of the proposition that "the live viewer is no more likely to buy prerecorded videotapes than is the time-shifter." There is not even a reference to the district court's findings. The

service, and the VCR as a *device*: those running services have much greater right and ability to control how their users deploy what they offer or sell.[43] This makes liability far more likely than in the case of the VCR.

Assuming infringement (either direct or indirect) is found, AI firms will then raise a fair use defense for the works generated by their systems. The fair use doctrine permits certain uses of copyrighted material that are unauthorized by the copyright owner.[44] Fair use defenses often boil down to highly contextual and contestable analyses of four factors,[45] which include:

> "(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."[46]

Courts may reject many AI providers' fair use defenses. It is unlikely that AI's generation of texts, images, sounds, and videos that are identical or substantially similar to copyrighted works would be held transformative under the first factor,[47] except in some exceptional cases

---

> majority should have left this point alone, or at least prefaced it with the more proper observation that the respondents failed to demonstrate via a preponderance of the evidence that time shifting does not dampen demand for prerecorded videotapes.

Frank Pasquale, Breaking the Vicious Circularity: *Sony*'s Contribution to the Fair Use Doctrine, 55 Case W. Res. L. Rev. 777, 793–94 n.65 (2005).

[43] Randall C. Picker, Rewinding *Sony*: The Evolving Product, Phoning Home and the Duty of Ongoing Design, 55 Case W. Res. L. Rev. 749, 759–61 (2005).

[44] Ruth Okediji, Givers, Takers, and Other Kinds of Users: A Fair Use Doctrine for Cyberspace, 53 Fla. L. Rev. 107, 117 (2001).

[45] See Barton Beebe, An Empirical Study of U.S. Copyright Fair Use Opinions, 1978-2005, 156 U. Pa. L. Rev. 549, 552-53 (2008); Haochen Sun, Copyright Law as an Engine of Public Interest Protection, 16 Nw. J. Tech. & Intell. Prop. 123, 124 (2019).

[46] 17 U.S.C. § 107.

[47] See Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 550 (2023) ("Goldsmith's original photograph of Prince, and AWF's copying use of that photograph in an image licensed to a special edition magazine devoted to Prince, share substantially the same purpose, and the use is of a commercial nature. AWF has offered no other persuasive justification for its unauthorized use of the photograph."). On the fourth factor, Fox News Network, LLC v. TVEyes, Inc., 883

like parody.[48] The fourth factor may weigh against fair use, given current and potential licensing arrangements.[49]

Training of AI is more likely to be treated favorably under fair use doctrine.[50] However, its legal treatment is by no means certain. As David Opderbeck puts it, while "[s]ome scholars and commentators argue that publicly accessible information should be available for AI training under a principle of non-expressive fair use," 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."[51] Opderbeck argues that "licensing regimes" for AI training data "would intersect productively with AI policy regarding fairness, transparency, privacy, and accountability."[52] Even scholars who believe that training AI should be a fair use of copyrighted work have acknowledged that important cases have "thrown the legality of machine copying [for purposes of machine learning] into question."[53]

For example, key aspects of training could be analogized to the copying of journal articles for Texaco scientists' research purposes,

F.3d 169, 180 ("In short, by selling access to Fox's audiovisual content without a license, TVEyes deprives Fox of revenues to which Fox is entitled as the copyright holder.")

[48] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994) ("[P]arody has an obvious claim to transformative value….").

[49] Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169, 180 ("[B]y selling access to Fox's audiovisual content without a license, TVEyes deprives Fox of revenues to which Fox is entitled as the copyright holder.").

[50] See Matthew Sag, The New Legal Landscape for Text Mining and Machine Learning, 66 J. Copyright Soc'y 291, 314–28 (2019) (applying the fair use factors to text data mining).

[51] David W. Opderbeck, Copyright in AI Training Data: A Human-Centered Approach, 76 Okla. L. Rev. (forthcoming 2024) (manuscript at 52), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4679299 [https://perma.cc/D9Q2-8M97].

[52] Id.

[53] Lemley & Casey, supra note 14, at 746. These cases include Am. Geophysical Union v. Texaco Inc., 60 F.3d 913, 914, 931 (2d Cir. 1994) (requiring Texaco to pay a licensing fee for internal copying of articles in academic journals), and Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169, 180–81 (2d Cir. 2018) (denying fair use defense for video search engine).

which was not held to be a fair use.[54] Established in 1977, the Copyright Clearance Center (CCC) has managed to arrange for licensing fees for varied texts utilized by both for-profit and non-profit entities, and its existence helped convince the *Texaco* court that copyright owners and users would be able to find mutually beneficial licensing deals to enable research.[55] The CCC followed in the footsteps of performing rights societies, like BMI and ASCAP, which have for decades arranged voluntary blanket licenses for the performance of copyrighted works.[56] In other situations, Congress has mandated a compulsory license for copyrighted works, bypassing questions of consent and simply requiring compensation in exchange for certain uses of works.[57]

Given that there is substantial uncertainty over the legality of AI providers' use of copyrighted works, legislators will need to articulate a bold new vision for rebalancing rights and responsibilities, just as they did in the wake of the development of the Internet (leading to the Digital Millennium Copyright Act of 1998). Parts III and IV below provide such a vision. To demonstrate its necessity, we first examine in Sections B and C below how untrammeled, unregulated use of copyrighted works by AI providers poses grave risks to creatives and to AI itself.

---

[54] Am. Geophysical Union, 60 F.3d at 930–31; see also Princeton Univ. Press v. Mich. Document Servs., 99 F.3d 1381, 1386–87 (6th Cir. 1997) (en banc) (holding that defendant's photocopying of plaintiff's copyrighted work was not a fair use because it harmed the reasonable potential market value of the copyrighted works).

[55] 60 F.3d at 930–31.

[56] I. Fred Koenigsberg, Performing Rights in Music and Performing Rights Organizations, Revisited, 50 J. Copyright Soc'y 355, 385–87 (2003).

[57] See, e.g., 17 U.S.C. § 111 (compulsory license for cable systems); 17 U.S.C. § 115 ("mechanical license" for making and distributing phonorecords); 17 U.S.C. § 119 (statutory license for satellite retransmissions for private home viewing). For a thoughtful examination of the potential for compulsory licensing to be more fair than blanket fair use determinations in scenarios involving new technological uses of copyrighted work, see generally Jacob Noti-Victor, Utility-Expanding Fair Use, 105 Minn. L. Rev. 1887 (2021) (describing how compulsory licensing can be adapted to suit technologies that make content accessible).

B. Market Substitution

AI presents a complex of threats to authors' livelihoods which are hard to analogize to past technologies. Photocopy machines simply copied past works. Cameras and past computers have depended on intense and extensive human supervision to create new works. Even though many authors complained about Google's copying of their works into the Google Books database, the database was ultimately a search tool, leading interested users to works that they could potentially buy.[58] It was not itself creating works. AI is different, as it is being promoted as general-purpose tool to create text, images, and audiovisual works at a rapid pace, with higher levels of quality expected over time. Pervasive secrecy also helps the firms avoid compensating the rights owners of copyrighted works like books, articles, music, images, and videos for their contributions.[59]

In July 2023, a U.S. Senate subcommittee discussed the licensing of AI training data. Senator Mazie Hirono questioned Stability AI's representative, Ben Brooks, about the company's position on paying for data used in training its AI models, with Brooks confirming that no payment arrangement was in place.[60] Technology companies have also failed to provide compensation for AI-generated works that are identical or substantially similar to authors' works.[61]

---

[58] Authors Guild v. Google, Inc., 804 F.3d 202, 207 (2d Cir. 2015) (finding Google's copying to be fair use); Frank Pasquale, Copyright in an Era of Information Overload: Toward the Privileging of Categorizers, 60 Vand. L. Rev. 133 (2007).

[59] See Karen Hao, We Don't Actually Know if AI is Taking Over Everything, The Atlantic (Oct. 19, 2023), https://www.theatlantic.com/technology/archive/2023/10/ai-technology-secrecy-transparency-index/675699 [https://perma.cc/U7T4-YRZK]; see also Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information 216–17 (2015) (describing the societal harms of pervasive secrecy).

[60] Jocelyn Noveck & Matt O'Brien, Visual Artists Fight Back Against AI Companies for Repurposing Their Work, Associated Press (Aug. 31, 2023, 2:55 PM), https://apnews.com/article/artists-ai-image-generators-stable-diffusion-midjourney-7ebcb6e6ddca3f165a3065c70ce85904 [https://perma.cc/WF7B-SLXW].

[61] See Gil Appel, Juliana Neelbauer & David A. Schweidel, Generative AI Has an Intellectual Property Problem, Harv. Bus. Rev. (Apr. 7, 2023),

The lack of compensation for authors' contributions to AI systems poses even more significant consequences for the future marketability of the authors' works. Firms and persons using AI systems have the potential to replace human authors by mimicking their writing style, mimicking important aspects of their work, or creating new content that is more desired or desirable, or better-marketed.[62]

AI systems can also rapidly produce vast amounts of content, making them an attractive option for organizations that require swift or high-volume content generation.[63] This could lead to an increased reliance on AI-generated news articles, marketing materials, and technical documentation, potentially reducing the demand for skilled practitioners in journalism, marketing, and technical writing.

Many creatives are alarmed by these trends. According to a survey undertaken by the Authors Guild in 2023, "69 percent of authors think their careers are threatened by [AI]," and "70 percent believe publishers will begin using AI to generate books in whole or part."[64] These concerns are not overstated. It is predicted that by 2025, 90% of content may be at least partially AI-driven.[65] In the realm of music, according to a 2023 survey, 73% of music producers have doubts about the security of their roles in the creative process, sensing the encroaching presence of AI.[66]

---

https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem [https://perma.cc/8K2J-F466].

[62] See Davenport & Mittal, supra note 9.

[63] Id.

[64] Survey Reveals 90 Percent of Writers Believe Authors Should Be Compensated for the Use of Their Books in Training Generative AI, The Authors Guild (May 15, 2023), https://authorsguild.org/news/ai-survey-90-percent-of-writers-believe-authors-should-be-compensated-for-ai-training-use [https://perma.cc/PR8N-JRGP].

[65] Carolyn Giardina, CES: Could 90 Percent of Content Be AI-Driven by 2025?, The Hollywood Rep. (Jan. 8, 2023, 12:11 PM), https://www.hollywoodreporter.com/movies/movie-news/ces-ai-sag-aftra-1235290431 [https://perma.cc/D4A7-V75K].

[66] Cameron Sunkel, Survey Finds 73% of Music Producers Believe Artificial Intelligence Will Replace Them, EDM (June 6, 2023), https://edm.com/gear-

C. The Danger of Model Collapse

Ironically, a policy of free appropriation of copyrighted work may even menace AI development itself. Simply put, it is not sustainable to expect training data to persist as a renewable resource when it is being mined, without compensation, in part to create substitutes for itself.[67] Scholars in the field have identified a danger of LLMs "learning from data produced by other models," a possibility that is more likely the less humans are compensated for their work.[68] The researchers call this pathological outcome "model collapse," "a degenerative process whereby, over time, models forget the true underlying data distribution, even in the absence of a shift in the distribution over time."[69] Consider, for instance, a distribution of articles about a given topic existing at Time 1. Over time, early LLMs may generate material based on those articles. As later LLMs at Time 2 take in both the original human content, and the later LLM-generated content, their results can be skewed by the earlier LLMs' random or otherwise unjustified selection and arrangement of key points from the

tech/survey-music-producers-believe-ai-will-replace-them [https://perma.cc/M4DY-E2MP].

[67] To be sure, some texts and images (such as emails and selfies) may accumulate rapidly without copyright protection, as they are primarily created due to needs and desires distinct from the incentives copyright can provide. Scientific research also has independent foundations for creation. However, there are many other areas where the creation of new content is heavily reliant on copyright-derived funding, or on the assurance that copyright ownership permits control of works. As to the latter point, artist Jingna Zhang puts it well: "Words can't describe how dehumanizing it is to see my name used 20,000+ times in MidJourney. My life's work and who I am—reduced to meaningless fodder for a commercial image slot machine." Jingna Zhang (@zemotion), X, (Mar. 9, 2024, 12:19 AM), https://twitter.com/zemotion/status/1766332997312057415 [https://perma.cc/6N6U-QDS2].

[68] See Ilia Shumailov et al., The Curse of Recursion: Training on Generated Data Makes Models Forget 2 (May 31, 2023) (unpublished manuscript), https://arxiv.org/pdf/2305.17493.pdf [https://perma.cc/ATC8-P77P].

[69] Id.

human content, as well as the well-documented problems of hallucination and fabrication by LLMs.[70]

LLMs are *language* models, not *knowledge* models, and have no ability to independently reason about what is in the human-generated articles or images they process. Nor is text generated in response to requests for fiction or creative non-fiction reflective of a mind capable of apprehending the world, since LLMs are mere text-predictors. They do not interact with and sense the world as humans do.[71] LLMs increasingly based on earlier LLM output may become, after sufficient iterations, like the faded analog copies of copies of copies that are familiar to those who recall widespread distribution of materials via copy machines—many of which became almost unrecognizably blurred and distorted over time.[72]

The bottom line here is grim. If uncompensated and uncontrolled expropriation of copyrighted works continues, many creatives are likely to be further demoralized and eventually defunded as AI unfairly outcompetes them, or effectively drowns them out. Low-cost automated content will strike many as a cornucopian gift—until it becomes clear that AI itself is dependent on ongoing input of human-generated works in order to improve and remain relevant in a changing world. At that point, it may be too late to reinvigorate creative industries left moribund by neglect. Much of an entire generation of writers, composers, journalists, actors, and other creatives may be

[70] See, e.g., Yue Zhang et al., Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models 3–6 (Sept. 24, 2023) (unpublished manuscript), https://arxiv.org/pdf/2309.01219.pdf [https://perma.cc/V8SK-KQBV] (describing the types of LLM hallucinations).

[71] Noam Chomsky, Ian Roberts & Jeffrey Watumull, Noam Chomsky: The False Promise of ChatGPT, N.Y. Times (Mar. 8, 2023), https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html [https://perma.cc/4G5H-59GD] ("The human mind is not, like ChatGPT and its ilk, a lumbering statistical engine . . . .").

[72] Ted Chiang, ChatGPT is a Blurry JPEG of the Web, New Yorker (Feb. 9, 2023), https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web [https://perma.cc/8WRB-HN8B].

missing, dissuaded from even trying to publish, disseminate, or profit from their expression, given how easily aspects of their expression can be mimicked via AI, and how rapidly their own contributions may be occluded or overwhelmed by AI expression. [73] Legislative interventions are critical to avoid such an unfair and ultimately self-defeating outcome. Part III below describes a new opt-out mechanism that would give creatives more say over how their works are used.

III. AN OPT-OUT MECHANISM FOR COPYRIGHT OWNERS

Regulators must reinstate creatives' proprietary control of their works within an equitable knowledge ecosystem. A new opt-out mechanism for copyright owners would empower them to reclaim proprietary control of their works through streamlined "notice and action" procedures aimed at AI providers. This mechanism would allow authors to submit requests to such providers for the removal of their works from the datasets of relevant AI systems, and take additional actions, as described in Sections A and B below.

For ease of reference, we will assume for the rest of this article that the authors of given works discussed also own the copyright in those works. To be sure, copyrights are often bought by publishers, recording companies, and other firms, or accrue to institutions in work-made-for-hire scenarios. In such cases, our proposal is agnostic as to whether authors, copyright owners, or both should be able to deny

[73] On the latter point, see Frank Pasquale, Cultural Foundations for Conserving Human Capacities in an Era of Generative Artificial Intelligence: Toward a Philosophico-Literary Critique of Simulation, *in* Being Human (forthcoming 2024) (manuscript at 1) (Beate Rossler & Valerie Steeves eds.) (on file with authors) ("Within a few years, machine-written language may become 'the norm and human-written prose the exception.' Generative AI is now poised to create profiles on social media sites and post far more than any human can—perhaps by orders of magnitude. Unscrupulous academics and public relations firms may use article-generating and -submitting AI to spam journals and journalists. The science fiction magazine *Clarkesworld* closed down its open submission window because of a deluge of LLM-written or assisted content." (citations omitted)).

consent to use their works in AI. This is a detail that would need to be worked out in a legislative process, perhaps with some reference to past debates on the proper extent and scope of moral rights, since many creatives will have ethical and cultural objections to works they created being used in AI training or reproduced by AI.

A. NOTICE AND ACTION PROCEDURES

Under the proposed mechanism, copyright owners can first request AI providers to take actions to effectively prevent their systems from generating outputs that appear identical or substantially similar to relevant copyrighted works. A copyright owner would be entitled to send a notice to an AI provider when he or she identifies that an output generated by the provider's AI system contains either a verbatim or substantially similar copy of his or her work, or a derivative work. In the notice, the copyright owner would be obliged to document the unauthorized reproduction of the work and his or her copyright ownership, along with a digital copy or an online link to the work.

The notice would target AI-generated content that resembles or adapts the copyright owner's work, potentially infringing upon the author's right to reproduction[74] or right to prepare derivative works.[75] For example, an exact replica of a copyrighted image or video generated by an AI system is highly likely to infringe on the right of reproduction. At the same time, certain adaptations of works by AI systems could also infringe on the right to prepare derivative works. [76] Such adaptations might include an image generator creating a painting based on a photograph, a chatbot condensing a novel into a novella,

[74] 17 U.S.C. § 106(1).
[75] Id. § 106(2).
[76] Daniel J. Gervais, AI Derivatives: The Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines, 52 Seton Hall L. Rev. 1111, 1112–13 (2022).

and a sound generator composing or performing a soundtrack derived from preexisting songs.

Upon receiving the notice, if the AI provider concerned accepts that the author has established a *prima facie* case of copyright infringement, it must promptly take actions to prevent such infringing content from being generated by its system again.[77] These actions may include: (1) removing the author's work from the datasets used by its AI system; (2) embedding filtering technology into the AI system to prevent generation of similar content that would infringe the author's copyright again; [78] and (3) initiating a "machine unlearning" process to remove the influence of the author's work from the AI system.[79] Upon completion, the AI provider should inform the author of the actions taken and provide an appropriate explanation of the effects of such actions.

Under the proposed mechanism, AI providers are obligated to take relevant actions expeditiously in response to notices submitted by

[77] Conversely, if the AI provider reasonably believes that the notice lacks valid legal grounds, it may send a counter-notification to the author, explaining reasons for not complying with the request. Further contestation procedures are beyond the scope of this essay, but may culminate in standard copyright litigation. The threat of enhanced statutory damages for willfulness will act as a strong deterrent to AI providers' ignoring or frivolously contesting valid complaints from copyright owners. 17 U.S.C. § 504(c)(1), 504(c)(2) (indicating that while standard statutory damages range between $750 and $30,000, in "a case where the copyright owner sustains the burden of proving, and the court finds, that infringement was committed willfully, the court in its discretion may increase the award of statutory damages to a sum of not more than $150,000.").

[78] Haochen Sun, The Ethics of AI Creativity 9 (Mar. 2024) (unpublished manuscript) (on file with authors) (proposing that "AI companies should be legally required to proactively implement filtering technologies that monitor and remove AI-generated works that appear identical or substantially similar to copyrighted works.").

[79] On machine unlearning, see Lucas Bourtoule et al., Machine Unlearning 1 (Dec. 15, 2020) (unpublished manuscript), https://arxiv.org/abs/1912.03817 [https://perma.cc/W3B8-TD9B]; Luciano Floridi, Machine Unlearning: Its Nature, Scope, and Importance for a "Delete Culture," Phil. & Tech., June 14, 2023, at 1, 2–4. Given that these actions may entail expensive model retraining, we envision an annual deadline for notification of AI providers by objecting authors, and another deadline for authoritative resolution of claims. Model retraining in response to copyright objections would then be no more than a yearly occurrence.

authors.[80] For more complex tasks, such as machine unlearning, providers should be granted additional time, as long as they take action in good faith. The proposed mechanism would initially impose monetary penalties on AI providers if they fail to promptly reply to legally valid notices submitted by copyright owners.

AI providers may take several actions to safeguard valid interests of copyright owners as requested through the notices. Regarding the first major action that AI providers can undertake, numerous copyright owners have requested the removal of their works from AI systems' datasets (or, when that is not possible, destroying the copy of the dataset including their works).[81] Going forward, technology can assist here: online platforms have already employed copyright filtering technology to detect infringing content and prevent it from being uploaded. Similarly, AI providers have developed and implemented filtering technologies, such as Microsoft's Copilot and OpenAI's Copyright Shield, to minimize instances of copyright infringement caused by their systems' generated content.[82] Some AI providers have devised innovative methods, enabling their models to selectively "unlearn" specific information. For example, by replacing particular content in the model's dataset with generic data, Microsoft researchers

[80] Courts will need to clarify the meaning of the term "expeditious," as they have done so in the context of DMCA notice and take-down cases. For relatively simple tasks, actions in that context have been considered expeditious if completed on the same day, or a few weeks after, the copyright owner sent a proper notice. See Capitol Records, LLC v. Vimeo, LLC, 972 F. Supp. 2d 500, 536 (S.D.N.Y. 2013); Wolk v. Kodak Imaging Network, Inc., 840 F. Supp. 2d 724, 733, 747 (S.D.N.Y. 2012).

[81] See, e.g., Complaint at 68, N.Y. Times Comp. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023) ("Ordering destruction under 17 U.S.C. § 503(b) of all GPT or other LLM models and training sets that incorporate Times Works . . . .").

[82] See, e.g., Brad Smith, 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/LRJ3-3F27]; Kyle Wiggers, OpenAI Promises to Defend Business Customers Against Copyright Claims, TechCrunch (Nov. 6, 2023, 1:15 PM), https://techcrunch.com/2023/11/06/openai-promises-to-defend-business-customers-against-copyright-claims/ [https://perma.cc/Y3HX-DYYL].

have successfully made the model forget details related to Harry Potter.[83]

Hence, copyright owners may complement their removal requests with demands that AI providers take additional actions. They could ask AI providers to adjust the operation of the filtering technology to prevent the generation of copyright-infringing content. If AI providers are able to develop and apply machine unlearning technology, copyright owners may request them to utilize it to make their AI models "forget" authors' works.

Though AI providers' data and methods are often secret, copyright owners have several options for detecting AI-generated content that infringes their copyrights. A straightforward method is for copyright owners to test an AI system themselves. For example, to determine whether an AI system allows users to create exact replicas, authors can input prompts such as "make an exact copy of X."[84] Copyright owners may also come across infringing content given marketing of AI system capabilities or programmed disclosures of provenance.[85] Furthermore, the application of watermarks to AI-generated content can also facilitate copyright owners' detection of infringing activities. Such watermarks may indicate the AI-generated nature of an output and the specific system that generated it.[86]

Processing notices from authors could initiate a dialogue between authors and AI providers, fostering discussions on the most

[83] See Ronen Eldan & Mark Russinovich, Who's Harry Potter? Approximate Unlearning in LLMs 2–3, 6–8 (Oct. 4, 2023) (unpublished manuscript), https://arxiv.org/abs/2310.02238 [https://perma.cc/E9UU-P4TS].

[84] João Pedro Quintais, Generative AI, Copyright and the AI Act, Kluwer Copyright Blog (May 9, 2023), https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act [https://perma.cc/66N3-ZFCK] (illustrating the image outputs of "exact copy" prompts).

[85] Haochen Sun, Redesigning Copyright Protection in the Era of Artificial Intelligence, 107 Iowa L. Rev. 1213, 1246 (2022) ("For now, this is straightforward because AI developers spontaneously announce this 'AI-generated works' status to publicize the development of their AI systems.").

[86] Sun, supra note 78, at 63–66 (suggesting that AI providers should be legally obligated to apply watermarks to show the AI-generated nature of the content).

effective methods to prevent copyright infringement. Considering the relative ease of determining whether two works are identical, it should be feasible for AI developers to process the first type of notices swiftly, since they pertain to exact replicas. However, it would often be more difficult or controversial to ascertain substantial similarity between two works, as this can involve a more complex evaluation of the elements shared by the original and the AI-generated content, as well as potential defenses (such as independent creation or fair use). In such cases, open communication and collaboration between authors and AI providers would become crucial to address concerns and find solutions that balance copyright protection with technological innovation.[87] When common agreement cannot be reached, traditional venues for litigation are available.

B. NORMATIVE RATIONALES FOR THE PROPOSED OPT-OUT MECHANISM

The proposed mechanism would effectively empower authors to opt out of AI systems that generate content infringing on their copyrights. As a result, it would enhance authors' control over their works, enabling them to better protect their interests amidst the surge of copyright infringement facilitated by AI systems. Even if direct copying of copyrighted works by AI systems occurs in a small percentage of cases, it may still have great impact given AI systems' vast output. One study of Stable Diffusion found that its models copied from its training data approximately 1.88% of the time.[88] Given that AI is estimated to have generated over 150 billion images in a single

[87] See Howard Hogan, Connor Sullivan & Jeffrey Myers, Copyright Liability for Generative AI Pivots on Fair Use Doctrine, Bloomberg Law (Sept. 22, 2023, 4:00 AM), https://news.bloomberglaw.com/us-law-week/copyright-liability-for-generative-ai-pivots-on-fair-use-doctrine [https://perma.cc/G53G-EEWU].

[88] Kyle Wiggers, Image-Generating AI Can Copy and Paste from Training Data, Raising IP Concerns, TechCrunch (Dec. 13, 2022, 7:30 AM), https://techcrunch.com/2022/12/13/image-generating-ai-can-copy-and-paste-from-training-data-raising-ip-concerns [https://perma.cc/3H2Q-99T7].

year,[89] even such a small percentage will generate myriad infringing images. The chance that the output of AI systems would infringe on at least some copyrighted content is high.[90] In response, the opt-out mechanism aims to minimize the impacts of infringing activities on authors' interests in multiple ways.

First, the proposed mechanism draws on methods developed in a long-standing copyright regime governing the use of content online, while adapting them to the age of AI. The DMCA established a robust notice and takedown process online, enabling authors to swiftly remove copyright-infringing content from platforms.[91] Copyright owners aggrieved by AI providers' infringing outputs have, at present, no recourse to such procedures, since AI creates works, rather than hosting and arranging them in the manner of DMCA-covered online service providers like YouTube or Facebook.[92] The DMCA offers protections to platforms that host infringing user-generated content, on the rationale that the platform cannot preemptively police users'

[89] Lea Zeitoun, AI Has Generated 150 Years Worth of Images in Less than 12 Months, Study Shows, Designboom (Aug. 21, 2023), https://www.designboom.com/technology/ai-has-generated-150-years-worth-of-photographs-in-less-than-12-months-study-shows-08-21-2023 [https://perma.cc/U8PJ-PJHR].

[90] For an example supporting this analysis, see Matthew Sag, Copyright Safety for Generative AI, 61 Hous. L. Rev. 295, 327–31 (2023) (concluding that copyrightable characters may easily provoke copyright infringement by AI systems).

[91] 17 U.S.C. § 512. In the case of the DMCA, Congress also afforded certain entities immunity from liability for copyright infringement if they abided by a number of conditions described in the Act. Id. § 512(b)–(c). The opt-out mechanism we propose could be coupled with a similar safe harbor. For example, Congress could grant AI providers a royalty-free statutory license to use copyrighted works in AI training until a copyright owner submits a valid objection. We do not take a position on the wisdom of this approach. It is one way to resolve current legal uncertainty over AI providers' use of copyrighted works. However, it does not seem to be as merited in the case of AI as it is in the case of, say, hosts of user-generated content, since the AI itself is the entity often generating (rather than merely hosting) the content.

[92] The DMCA notice-and-takedown regime does not extend to situations where infringing works are generated by AI systems at the behest of their users. Peter Henderson et al., Foundation Models and Fair Use 18 (Mar. 29, 2023), https://arxiv.org/pdf/2303.15715.pdf [https://perma.cc/Q7MK-7K6G] (observing that "generated content does not have the same safe harbor and that post-hoc take-downs are not sufficient to reduce liability").

actions.[93] AI content, by contrast, is being created by the AI firm itself, so it is clearly responsible for it under the proposed mechanism.

Second, the mechanism would serve an information-forcing function, empowering copyright owners to address infringing activities perpetrated by opaque sociotechnical systems utilizing AI. Copyright owners need a new mechanism allowing them to compel AI providers to disclose information about how their works are used in training models and generating content. Any AI provider wishing to contest a notice from the author would need to provide explanations about its methods, such as whether its datasets contain the copyrighted work in question, and the workings of technologies concerning content removal, filtering, and unlearning. Thus, this new mechanism would enable authors to regain proprietary control over how their works are used in in AI systems.

Third, the proposed mechanism would also address major problems with existing opt-out procedures offered by some AI providers. For example, OpenAI purports to provide creatives with an option to avoid incorporating their creations among the photos, paintings, and other visual items that its AI systems, such as DALL-E, utilize for training and subsequent image generation. However, many creatives claim that AI providers' opt-out processes are burdensome and complex.[94] Authors have lamented that such opt-out procedures are "a bad joke" and "a fake PR stunt" for AI providers.[95] Self-regulation will not be effective here.

Last but not least, the proposed mechanism would also provide authors with a more efficient and cost-effective alternative for dispute

[93] Fromer & Sprigman, supra note 19, at 527.
[94] Matteo Wong, Artists Are Losing the War Against AI, The Atlantic (Oct. 2, 2023), https://www.theatlantic.com/technology/archive/2023/10/openai-dall-e-3-artists-work/675519 [https://perma.cc/8J92-WE8B].
[95] Kate Knibbs, Artists Allege Meta's AI Data Deletion Request Process Is a "Fake PR Stunt," Wired (Oct. 26, 2023, 7:00 AM), https://www.wired.com/story/meta-artificial-intelligence-data-deletion [https://perma.cc/R72V-EKF4].

resolution than the judicial process. Litigation is often time-consuming and expensive. According to the American Intellectual Property Law Association, litigating a single copyright infringement case in a U.S. federal court from pre-trial to appeals costs an average of $278,000 and may take over a year in many instances.[96] Such litigation costs might not pose a problem for large corporations and the wealthiest creatives and content owners. For example, when *The New York Times* credibly threatened to sue OpenAI for using its content for data training without consent, Common Crawl removed links to *The New York Times*'s content from its datasets.[97] However, *The New York Times* still felt obliged to sue both OpenAI and Microsoft a few months later.[98] Many authors lack the financial resources to litigate against AI vendors (many of which are massive firms, or are backed by such firms) over potentially lengthy periods.

In contrast, an opt-out mechanism offers authors a streamlined and cost-effective way to assert their rights. AI providers would be required to promptly review the request from an author, make a decision, and notify the complainant of their decision. Failure to do so in a good faith manner should subject the firm to civil penalties. Hence, this approach would ensure that authors have an accessible and

[96] A Guide to Intellectual Property Litigation, Thomson Reuters (Dec. 23, 2022), https://legal.thomsonreuters.com/blog/guide-to-intellectual-property-litigation [https://perma.cc/W6CT-DG63].

[97] Bobby Allyn, "New York Times" Considers Legal Action Against OpenAI as Copyright Tensions Swirl, NPR (Aug. 16, 2023, 5:53 PM), https://www.npr.org/2023/08/16/1194202562/new-york-times-considers-legal-action-against-openai-as-copyright-tensions-swirl [https://perma.cc/5HGD-5ZSN]; Alistair Barr & Kali Hays, The New York Times Got Its Content Removed from One of the Biggest AI Training Datasets. Here's How It Did It., Bus. Insider (Nov. 8, 2023, 5:00 AM), https://www.businessinsider.com/new-york-times-content-removed-common-crawl-ai-training-dataset-2023-11 [https://perma.cc/657G-BHSK].

[98] Clare Duffy & David Goldman, The New York Times Sues OpenAI and Microsoft for Copyright Infringement, CNN, https://edition.cnn.com/2023/12/27/tech/new-york-times-sues-openai-microsoft/index.html [https://perma.cc/E9ZB-Z4H8] (last updated Dec. 27, 2023, 6:02 PM).

efficient means of protecting their works before resorting to potentially lengthy and costly litigation.

It is important to emphasize that the proposed mechanism is not designed to enable creatives to undermine the fair use privileges that may be enjoyed by AI providers. The fair use doctrine generally does not authorize the creation of new works that infringe on another's copyright, such as by making an exact copy without transformative use.[99] Moreover, the proposed mechanism does not grant authors the right to prevent AI providers from using their works for data training processes without first documenting copyright infringement arising out of content generation. When an AI provider has a good faith belief that the output it generates is a fair use of the copyrighted work, it simply needs to indicate the basis of that belief in a reply to a complaining copyright owner, to avoid the civil penalties mentioned above.

To be sure, the proposed opt-out mechanism is no panacea. If it is used too widely, it may corrode the quality of training data. Consider the larger social implications of *The New York Times*'s departure from AI datasets if, for example, the newspaper successfully opted out. This would leave a significant hole in journalistic data, given the quality of *The New York Times*'s coverage and its exacting editorial standards.[100] Meanwhile, other outlets may fill the vacuum with biased or lower quality reporting.[101] LLM-generated news may also become more

[99] See Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1262 (2023) ("As most copying has some further purpose and many secondary works add something new, the first factor asks 'whether *and to what extent*' the use at issue has a purpose or character different from the original . . . . The larger the difference, the more likely the first factor weighs in favor of fair use.") (citing Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994)).

[100] In its pending lawsuit against OpenAI, *The New York Times* alleges that, "by OpenAI's own admission, high-quality content, including content from The Times, was more important and valuable for training the GPT models as compared to content taken from other, lower-quality sources." Complaint at 27, N.Y. Times Comp. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023).

[101] Kate Knibbs, Most Top News Sites Block AI Bots. Right-Wing Media Welcomes Them, Wired (Jan. 24, 2024, 7:00 AM), https://www.wired.com/story/most-news-sites-block-ai-bots-right-wing-media-welcomes-them [https://perma.cc/7RMX-YPKV].

prominent in future LLMs' training datasets, exacerbating the problem of model collapse described in Part II above. Given such concerns, we believe that it would be advisable to incentivize content owners to allow their works to be used in LLM training by offering compensation—a concern addressed in Part IV below.

IV. PROVIDING COMPENSATION FOR COPYRIGHT OWNERS

Ninety percent of authors in one recent survey believed that they should be compensated for their works' use in training AI.[102] Meanwhile, many AI providers appear to believe they owe nothing to creatives. This deep divide in expectations and attitudes may impede voluntary licensing deals between copyright owners and AI providers, which have so far focused on journalistic content. This necessitates exploration of alternative paths to compensation, described in more detail below.

This Part explores two dimensions of the controversy over compensation for the use and production of works via AI.[103] First, Section A addresses the "why" of compensation: how varied normative perspectives vindicate some level of payment to the copyright owners whose work is at the foundation of AI. Then, Section B addresses the "how" of compensation, surveying potentially instructive precedents

[102] Survey Reveals 90 Percent of Writers Believe Authors Should Be Compensated for the Use of Their Books in Training Generative AI, supra note 64.

[103] While past works in this vein have focused on the production phase of GenAI, this part is focused on compensation due for training. For examples of this past work, see Martin Senftleben, Generative AI and Author Remuneration, 54 Int'l Rev. Intell. Prop. & Competition L. 1535, 1537 (2023) (acknowledging that "remuneration could be made mandatory at the AI training stage," but concluding that "a legislative approach that focuses on the output/substitution dimension and seeks to introduce a lump-sum AI levy system is more promising than taking input and training activities as a reference point for remuneration"). Nevertheless, if training is ultimately determined to be a fair use, the ideas for compensation here could be useful in determining the proper compensation to be arranged in a legislative or judicial settlement of what are sure to be numerous lawsuits based on the works produced by AI systems.

for fixed payments or proportional revenue sharing for copyright owners.

A. NORMATIVE RATIONALES FOR COMPENSATION

For many turn-of-the-millennium advocates of an open Internet, copyright was a menace, constantly threatening to stifle innovation. By contrast, many artists and activists now see it as one of the few tools left to demand accountability from an extraordinarily concentrated and powerful technology industry. Several rationales explain this shift.

One important rationale is an evolving reframing of key normative foundations of intellectual property policy, from "open vs. closed" to "labor vs. capital."[104] Relaxing copyright may seem like a deregulatory path to open innovation, but the term "open" itself has been overused and in many ways misused. As one insightful paper recently observed:

> [S]ome companies have moved to embrace 'open' AI as a mechanism to entrench dominance, using the rhetoric of 'open' AI to expand market power, and investing in 'open' AI efforts in ways that allow them to set standards of development while benefiting from the free labor of open source contributors.[105]

[104] For explorations of labor framing in intellectual property scholarship, see Xiyin Tang, Intellectual Property Law as Labor Policy 6–7, 42 (Mar. 2024) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4761809 [https://perma.cc/6FUQ-TCVW]; Frank Pasquale, Joining or Changing the Conversation? Catholic Social Thought and Intellectual Property, 29 Cardozo Arts & Ent. L.J. 681, 722 (2011).

[105] David Gray Widder, Meredith Whittaker & Sarah Myers West, Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI 3 (Aug. 16, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807 [https://perma.cc/TN8B-KNMX].

Massive technology firms have become rich in part based on uncompensated, or under-compensated, contributions from both users and content providers.[106] Thus the narrative of the copyleft, which argues that big content owners exploit users, must be supplemented by another story: big technology firms exploiting labor without adequate (and, often, any) compensation. A #CreateDontScrape movement has capitalized on this sentiment, adopting the copyleft's rhetoric of distributional justice and democratization toward a very different end.[107]

#CreateDontScrape faces an uphill battle. AI firms' extraordinary wealth leaves them well-positioned to fight labor, environmental, and intellectual property standards.[108] Under-compensation is also endemic in the industry. Consider, for instance, the extraordinary exploitation of certain content moderators working at a firm used by a vendor of AI to moderate the content its models trained on. The content moderators said they were paid less than $1 an hour, and frequently encountered deeply disturbing content.[109] Whereas garment workers have recently won some victories at the state level to put a wage floor on their piecework, it is likely to be much more difficult for AI workers to gain similar rights, in part because of the wealth and power of their employers.[110] To be sure, content creators

[106] See Haochen Sun, Technology and the Public Interest 124 (2022) (pointing out that users contribute "content that is quantitatively and qualitatively essential to the rapid development and success of social media platforms").
[107] See, e.g., Jon Lam #CreateDontScrape (@JonLamArt), X (Mar. 5, 2023, 11:32 AM), https://twitter.com/JonLamArt/status/1632418770949148673 [https://perma.cc/TCX7-H8NJ].
[108] See Brendan Bordelon, Key Congress Staffers in AI Debate Are Funded by Tech Giants like Google and Microsoft, Politico (Dec. 3, 2023, 7:00 AM), https://www.politico.com/news/2023/12/03/congress-ai-fellows-tech-companies-00129701 [https://perma.cc/W88V-K5AT].
[109] Alex Kantrowitz, He Helped Train ChatGPT. It Traumatized Him, CMSWire (May 23, 2023), https://www.cmswire.com/digital-experience/he-helped-train-chatgpt-it-traumatized-him [https://perma.cc/3XWB-YSPP].
[110] Izzie Ramirez, It's Time to Break up with Fast Fashion, Vox (Nov. 14, 2023, 6:00 AM), https://www.vox.com/even-better/2023/11/14/23955673/fast-fashion-shein-hauls-environment-human-rights-violations [https://perma.cc/9CTN-PU8K].

are not subject to the type of direct exploitation suffered by sweatshop workers. However, it is difficult to deny that under-compensation for labor is all too prevalent a reality in the contemporary technology space. The real ethical dilemma here may be less open versus closed systems than an intensifying conflict between labor and capital, with the latter unjustly enriched by the former's work, and AI threatening to accelerate that upward redistribution of wealth.

Copyright doctrine also favors creation by humans, denying copyrightability to many wholly or mainly-AI-generated works.[111] The U.S. Copyright Office's repeated refusal to register many AI-generated works indicates a strong commitment to the primacy of human action with respect to the types of works that copyright is meant to promote.[112] Given the prospect of AI-generated works overwhelming human-created works without some legal rebalancing of rights and interests, this is yet another rationale for human-centric compensation.[113]

Critics of compensation schemes for authors will likely insist that the amount of money available for compensation will be insignificant once divided among rights holders whose work has been integrated into training data.[114] Some hypothetical valuations are

[111] Sun, supra note 85, at 1227.

[112] Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 88 Fed. Reg. 16190, 16192 (Mar. 16, 2023) (to be codified at 37 C.F.R. pt. 202).

[113] For an exploration of this "drowning out" effect, see Matthew Kirschenbaum, Prepare for the Textpocalypse, The Atlantic (Mar. 8, 2023), https://www.theatlantic.com/technology/archive/2023/03/ai-chatgpt-writing-language-models/673318/ [https://perma.cc/4XPS-82QM] ("[L]ast June, a tweaked version of GPT-J, an open-source model, was patched into the anonymous message board 4chan and posted 15,000 largely toxic messages in 24 hours . . . . What if . . . millions or billions of such posts every single day [began] flooding the open internet, commingling with search results, spreading across social-media platforms, infiltrating Wikipedia entries, and, above all, providing fodder to be mined for future generations of machine-learning systems? . . . We may quickly find ourselves facing a textpocalypse, where machine-written language becomes the norm and human-written prose the exception.").

[114] Pamela Samuelson, Fair Use Defenses in Disruptive Technology Cases, UCLA L. Rev. (forthcoming 2024) (manuscript at 79) ("The amounts paid to individual copyright owners would likely be very modest, and would be unlikely to provide significant financial support to authors and artists."),

useful to formulate a response here. A writer/programmer has estimated that there are nearly 200,000 e-books in the Books3 database used by one publicly released version of ChatGPT's services.[115] Assume that a small, one-time levy on OpenAI of $5 million were set aside to pay relevant book authors for their inclusion in Books3. That would amount to at least $25 per book, if it were divided evenly. This is a small amount, but it is not trivial. Alternatively, the firm now entitled to up to 49% of the profits of a subsidiary of OpenAI, Microsoft, could pay a $50 million levy with less than one-thousandth of its 2023 net income of over $70 billion.[116] That would amount to at least $250 per book under the assumptions mentioned above. Moreover, OpenAI and Microsoft are only two entities in the AI space, and the levy would be imposed on a whole category of companies. It might also be imposed annually, instead of just one time. As industry revenues grow, the levy could grow as well if it were set as a percentage of revenue, instead of a fee for use of a given dataset.[117] All these factors counter fears that levy funds would not be significant once divided among potential claimants.

To be sure, when levied funds are distributed, there will be tensions between normative commitments to ease of administrability and particularized recognition of merit. A principle of equal allocation

---

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4631726 [https://perma.cc/62TT-8ZF6].

[115] Alex Reisner, Revealed: The Authors Whose Pirated Books Are Powering Generative AI, The Atlantic (Sept. 25, 2023, 1:40 PM), https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063 [https://perma.cc/N8C8-2JC2].

[116] Tim Bradshaw et al., How Microsoft's multibillion-dollar alliance with OpenAI really works, Fin. Times, Dec. 15, 2023, at https://www.ft.com/content/458b162d-c97a-4464-8afc-72d65afb28ed ("Microsoft's billions [of dollars of investment into OpenAI]--which include huge investments in data centre infrastructure as OpenAI's 'exclusive cloud provider'--entitle it to up to 49 per cent of the profit generated by a subsidiary of OpenAI, according to people familiar with the deal."). The net revenue figure comes from Lionel Sujay Vailshery, Microsoft's net income from 2002 to 2023, Statista, https://www.statista.com/statistics/267808/net-income-of-microsoft-since-2002/ (Mar. 6, 2024).

[117] For example, the Audio Home Recording Act imposed a 2% royalty payment on certain digital audio recording devices. 17 U.S.C. § 1004(a)(1).

per work, within certain bounded categories, would advance administrability. However, some copyright owners are likely to demand special solicitude toward works that are particularly lengthy, well-structured, and authoritative—such as books from reputable publishers. Reconciling these competing commitments will take a fair amount of diplomacy, similar to past negotiations for legislative compromises designed to allocate government-collected funds fairly.

But let us suppose, for the sake of argument, that there is *no* fair way of allocating whatever funds are gathered via a levy on AI firms' use of works, and that the amount collected does not significantly increase the income of most copyright owners. There are, nevertheless, independent normative grounds for requiring some form of wealth transfer away from the AI firms expropriating copyrighted works. Consider the analogous realm of class action litigation, where deterrence-based theories have often been embraced. Brian Fitzpatrick has argued that a "purely deterrence-based theory of civil litigation might be indifferent between defendants paying those they have injured and defendants paying completely unrelated third parties."[118] The key point is to ensure that an entity that has committed a wrong loses at least some of the utility attributable to the wrong it committed.

The normative rationale for a levy on unlicensed use of works is even stronger when levy funds are directed toward those whose works have been used without consent. Unjust enrichment is "a very broad and flexible equitable doctrine, based on the principle that it is contrary to equity and good conscience for the defendant to retain a

[118] Brian T. Fitzpatrick, Do Class Action Lawyers Make Too Little?, 158 U. Pa. L. Rev. 2043, 2060 (2010). Even in settlements where class action defendants do not compensate victims, or compensate them very little, they often must make payments to plaintiffs' attorneys or non-profits committed to identifying and deterring future wrongdoing.

benefit that has come to [them] at the expense of the plaintiff."[119] It is based on several normative rationales relevant here.

One moral foundation of unjust enrichment claims is avoiding windfalls attributable to the property or services of another.[120] Ying Hu has already applied an unjust enrichment framework to unauthorized personal data collection by AI firms, describing "situations in which [AI] companies might be required to disgorge profits from the unlawful collection or use of personal data."[121] The same logic could apply *a fortiori* in a copyright context, where the results of persons' own labor, rather than observations and inferences about them most often made by others, are at stake.

Another rationale for redistribution related to the avoidance of unjust enrichment is the reduction of already vast power differentials. Legislators may also decide to reduce the ability of parties that have unfairly reaped benefits from another's labor from further leveraging this money into the power and influence to continue such oppression. Given the extraordinary power of the technology industry to fund think tanks, universities, and academics favorable to its ideology, and to block legislation adverse to its interests, a growing and self-reinforcing power asymmetry is a clear and present danger here.[122] Any level of transfer between leading AI providers and creatives would tend to abate it, regardless of its size or ultimate destination.

---

[119] SEC v. Sanchez-Diaz Monge, 88 F.4th 81, 88 (1st Cir. 2023) (citing George E. Palmer, Law of Restitution § 1.1 (3d ed. 2023)).

[120] Ayelet Gordon-Tapiero & Yotam Kaplan, Unjust Enrichment by Algorithm, Geo. Wash. L. Rev. (forthcoming 2024) (manuscript at 26), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4445833 [https://perma.cc/Q8YG-Y89G].

[121] Ying Hu, Unjust Enrichment Law and AI, *in* The Cambridge Handbook of Private Law and Artificial Intelligence 287, 288 (Ernest Lim & Phillip Morgan eds., 2024).

[122] See, e.g., Sun, supra note 104, at 121 ("[The major tech firms] have been the beneficiaries of lax statutory and regulatory arrangements, and are today among the most financially and politically powerful in the world."); Rebecca Klar & Karl Evers-Hillstrom, How Big Tech Fought Antitrust Reform—And Won, The Hill (Dec. 23, 2022, 6:00 AM), https://thehill.com/policy/technology/3785894-how-big-tech-fought-antitrust-reform-and-won [https://perma.cc/Z4UC-MGSY] (describing technology companies' lobbying efforts).

Nevertheless, there is a burden on proponents of a compensation scheme to estimate how substantial it should be. The next section articulates principles for such calculations, drawing on precedents in both creative industries and in other sectors of the economy. Levies, such as those imposed on certain digital audio recording devices under the Audio Home Recording Act (AHRA) of 1992, may raise funds to be distributed to copyright owners.[123] Scholars have already recommended imposing similar levies in the AI context.[124] These encouraging precedents could develop into administered pricing for use of the works of copyright owners who have not exercised the opt-out we proposed in Part III above.

B. BENCHMARKING COPYRIGHT COMPENSATION FOR GENERATIVE AI

Compensation has been an under-theorized aspect of the list of demands on AI providers made by creatives.[125] The question of the proper level of compensation for the copyright-protected texts, images, films, and other inputs used for training AI models is a difficult one, but valuation problems are far from insurmountable. The U.S. government already sets prices for many uses of music.[126] Much more complex and higher stakes economic arrangements have been subject to multiple forms of administered pricing.[127] Given these precedents,

[123] 17 U.S.C. § 1004(a)(1).

[124] See, e.g., Benjamin L.W. Sobel, Artificial Intelligence's Fair Use Crisis, 41 Colum. J.L. & Arts 45, 93 (2017); Christophe Geiger & Vincenzo Iaia, Comment, The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI, 52 Computer L. & Sec. Rev. 105925, 6 (2024).

[125] This section accepts Benjamin Sobel's prescient invitation to offer methods of estimating an appropriate level of compensation for use of copyrighted works in training AI models. Sobel, supra note 124, at 92 ("Calculating the appropriate levy, and prescribing its disbursement, is a more ambitious task than this Article can fulfill."). We leave for future work recommendations for the proper disbursement of the levy.

[126] U.S. Copyright Off., Copyright and the Music Marketplace: A Report of the Register of Copyrights 145 (2015).

[127] For example, as of 2022, the Medicare and Medicaid programs in the U.S. together administer almost 7% of gross domestic product via a highly complex mixture of direct payments, performance-based payments, and other subventions. Ctrs. for

and the complex administration of economic value and valuation in transport, communications, and other infrastructure in many jurisdictions, legislators should not shrink from working out compensation schemes here.[128]

A levy on AI providers using copyrighted works is one way to generate funds for compensation of affected copyright owners. The AHRA provides one precedent. The AHRA imposed a levy on sales of recording devices and media, anticipating their use in uncompensated and unauthorized copying of copyrighted work.[129] As with AI in the present, the rise of such devices in the past was seen as posing "threats to the livelihood of creative individuals and current or future copyright owners."[130] Not only sales, but also importation and distribution of devices, triggered the levy.[131] The default minimum levy for recording devices was 2% of the product's wholesale price, or $1, whichever was higher.[132] The maximum levy was $8 for a recording device.[133] Media faced a levy of 3% of the wholesale price, with no minimum or maximum level.[134] The funds collected were later distributed to artists,

Medicare & Medicaid Servs., National Health Expenditures Fact Sheet, https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet [https://perma.cc/Y5RK-QU6D] (last updated Dec. 13, 2023, 4:13 PM).

[128] For a recent text examining such valuation practices, see generally Morgan Ricks, Ganesh Sitaraman, Shelley Welton & Lev Menand, Networks, Platforms, and Utilities: Law and Policy (2022) (describing rate setting and valuation across industries).

[129] 17 U.S.C. §§ 1003–04.

[130] Joel L. McKuin, Home Audio Taping of Copyrighted Works and the Audio Home Recording Act of 1992: A Critical Analysis, 16 Hastings Comm. & Ent. L.J. 311, 313 (1994).

[131] 17 U.S.C. § 1003(a).

[132] 17 U.S.C. § 1004(a)(1), (3); see also Geoffrey Hull, The Home Recording Act of 1992: A Digital Dead Duck or Finally Coming Home to Roost?, 2 Music & Ent. Indust. Educators Ass'n J. 76 (2002) (describing the royalty system established under the AHRA).

[133] 17 U.S.C. § 1004(a)(3).

[134] 17 U.S.C. § 1004(b).

publishers, and related parties.[135] Japan, Canada, and the Netherlands have implemented similar levies.[136]

Given the complexity of the AI supply chain, particularly with respect to generative AI, it is not feasible to impose a per-device cost on AI providers. However, other triggers for payment are possible. Levies on the use of particular datasets may be imposed, or on model training, or on some aggregate number of responses provided to users, or on paid subscriptions. Alternatively, the level of the levy could be benchmarked with respect to some percentage of AI providers' expenditures or revenues.

Consider first the expenditure side. Leading AI providers depend on three critical inputs: expert personnel, advanced computing equipment, and massive amounts of training data. Top firms spend lavishly on the first two factors of production. Compensation for engineers has exceeded $800,000 per year at OpenAI;[137] top talent commands millions per year in salary and may ultimately earn tens or hundreds of millions of dollars via stock options or other equity-driven compensation. Moreover, firms selling computing equipment have become among the most valuable corporations globally. For example, Nvidia's market valuation exceeded $1 trillion in 2023, ranking it among the top ten most valuable firms in the world.[138] This valuation

[135] 17 U.S.C. § 1006(a).

[136] Salil K. Mehra, The iPod Tax: Why the Digital Copyright System of American Law Professors' Dreams Failed in Japan, 79 U. Colo. L. Rev. 421, 446–47, 463–64 (2008) (describing Japan's levy and recommending ways of improving future levies); Copyright Act R.S.C. 1985, c C-42 § 82 (describing Canadian requirements); Gov't of the Netherlands, What is the Private Copy Levy?, https://www.government.nl/topics/intellectual-property/question-and-answer/what-is-the-private-copy-levy [https://perma.cc/CQ5V-CMC4] (last visited Apr. 19, 2024) (describing the Netherlands's approach); Monica Zhang , 'Fair Compensation' in the Digital Age: Realigning the Audio Home Recording Act, 38 Hastings Comm. & Ent L.J. 145, 160–64 (2016) (surveying approaches).

[137] See Jo Constantz, OpenAI Engineers Earning $800,000 a Year Turn Rare Skillset into Leverage, Yahoo! Fin. (Nov. 22, 2023), https://finance.yahoo.com/news/openai-engineers-earning-800-000-183139353.html [https://perma.cc/8DCJ-Z9WN].

[138] Patturaja Murugaboopathy & Gaurav Dogra, Nvidia's Market Cap Climbs Amid Tech Turbulence in August, Reuters (Sept. 1, 2023, 8:03 AM),

was premised on revenues estimated at over $60 billion per year, and a significant share of those revenues is based on sales to AI providers.[139]

Based on a simple tripartite division, policymakers might conclude that training data, in the aggregate, is worth at least as much as either the computing talent or the infrastructure now used to process it. On this model, policymakers may impose a levy at a level meant to promote such parity. If policymakers found that excessive, a smaller percentage of firm spending could be earmarked for compensation for copyright owners.

Another way of calculating value would be premised on valuation of the services provided by firms providing AI—i.e., the revenue as opposed to the spending side of the equation. For example, a for-profit firm that makes $10 billion in revenue yearly may be required to allocate 5% of its revenues to a levy to be distributed to copyright owners who have not agreed to alternative licensing arrangements. A similar proposal recently reshaped debates on online advertising markets, by estimating that Google and Facebook would owe at least $11.9 billion annually to news providers in the U.S. if advertising revenue were split evenly between these platforms and content creators whose work provides so much of the platforms' value.[140] If a government considered imposing such revenue sharing, Facebook and Google would likely dispute the reasoning in the report and provide their own rationales for why news was worth less. Further

https://www.reuters.com/business/global-markets-marketcap-2023-09-01 [https://perma.cc/7F78-XL9K].

[139] NVIDIA Revenue 2010–2024, Macro Trends, https://www.macrotrends.net/stocks/charts/NVDA/nvidia/revenue [https://perma.cc/PY5G-WRT2] (last visited Mar. 6, 2024); Daniel Howley, Nvidia Stock Surges After Earnings Beat Estimates Across the Board (Feb. 22, 2024), https://finance.yahoo.com/news/nvidia-stock-surges-after-earnings-beat-estimates-across-the-board-161450767.html [https://perma.cc/8M95-6P44].

[140] Patrick Holder, Haaris Mateen, Anya Schiffrin & Haris Tabakovic, Paying for News: What Google and Meta Owe U.S. Publishers 4 (Nov. 13, 2023), https://policydialogue.org/files/publications/papers/LatestVersion.pdf [https://perma.cc/DRD3-VGUH].

replies and counter-replies would ensue. Out of such disputation, policymakers will eventually be in a position to make a reasoned determination about the proper level of compensation due. A similar dynamic could inform levies in the AI space.

Other jurisdictions have recently catalyzed public conversations about the economic relationship between technology firms and the media producers which bring so many profit-generating users and advertisers to them. Recognizing the need for a rebalancing of bargaining power online, Australia and Canada have enacted negotiation mechanisms for determining how much large search and social intermediaries owe to media and news organizations.[141] Several licensing deals have been struck in Australia.[142] OpenAI has itself recognized the value of news, at least, as it has licensed content from the *Financial Times*, the "US-based Associated Press, Germany's Axel Springer, France's Le Monde and Spain's Prisa Media."[143] It and other AI providers should acknowledge the fairness of extending such subventions, via either licensing or distribution of proceeds from a levy, to a broader set of sources.

[141] News Media Bargaining Code, Austl. Competition & Consumer Comm'n, https://www.accc.gov.au/by-industry/digital-platforms-and-services/news-media-bargaining-code/news-media-bargaining-code [https://perma.cc/NW52-LS3E] (last visited Mar. 5, 2024); The *Online News Act*, Gov't of Can. (Jan. 3, 2024), https://www.canada.ca/en/canadian-heritage/services/online-news.html [https://perma.cc/MM4Y-V4EB]. For more on the promise of negotiation in the context of AI, see Geiger & Iaia, supra note 124, at 8 ("[Q]uantification of remuneration rates can be left to negotiation.").

[142] News Media Bargaining Code, Austl. Competition & Consumer Comm'n, https://www.accc.gov.au/by-industry/digital-platforms-and-services/news-media-bargaining-code/news-media-bargaining-code [https://perma.cc/NW52-LS3E] (last visited May 6, 2024) (concluding that "the code has been a success to date. Over 30 commercial agreements between digital platforms (Google and Meta) and a cross section of Australian news businesses have been struck, agreements that were highly unlikely to have been made without the code.").

[143] Madhumita Murgia, The Financial Times and OpenAI strike content licensing deal, Fin. Times, Apr. 29, 2024.

A legislative solution should also be calibrated to the varied uses and purposes of AI.[144] An AI provider whose primary customers are writers who wish to use AI to generate poetry, and who have under $5,000 in sales, may fairly be expected to be levied very little or nothing. Similarly, non-profit, research-focused institutions like universities may properly face zero or small levies for their provision of AI. By contrast, a "news service" that provides AI in order to rewrite journalists' work into a composite (if technically non-infringing) story, undercutting the entities that actually invested in the journalism necessary to report, select, and arrange the underlying facts, should be required to pay an amount commensurate with the sums invested by the entities whose work it has copied.[145] This valuation goes to the social purpose of compensation—ensuring long-term production of knowledge, rather than ruinous competition and new forms of piracy that essentially make it impossible for any person or entity to invest in affected industries.

As compensation schemes are elaborated, they may distinguish between high-revenue and low-revenue entities. A detailed statute may carefully tailor proper levels for a levy, much as prior compulsory licenses or levies in copyright law have done.[146] Even private ordering, via blanket licenses like ASCAP's, tends to recognize this principle.[147]

---

[144] On the importance of such calibration, see Michael Carroll, One for All: The Problem of Uniformity Cost in Intellectual Property Law, 55 Am. U. L. Rev. 845, 852–55 (2006).

[145] This is a particularly important form of redress to pursue in the U.S. given federal preemption of hot news misappropriation claims. See, e.g., Barclays Capital Inc. v. Theflyonthewall.com, Inc., 650 F.3d 876, 907 (2d Cir. 2011) ("[A] Firm's ability to make news—by issuing a Recommendation that is likely to affect the market price of a security—does not give rise to a right for it to control who breaks that news and how.").

[146] Jacob Noti-Victor, Copyright's Law of Dissemination, 44 Cardozo L. Rev. 1769, 1789–98 (2023) (discussing nuanced compensation systems); Jacob Victor, Reconceptualizing Compulsory Copyright Licenses, 72 Stan. L. Rev. 915, 938–47 (2020) (discussing the history of U.S. statutory licensing regimes for music).

[147] Michael B. Rutner, Note, The ASCAP Licensing Model and the Internet: A Potential Solution to High-Tech Copyright Infringement, 39 B.C. L. Rev. 1061, 1076–78 (1998).

Accordingly, a small, non-profit entity may be permitted to enjoy free use of materials for training its models, or some subset of them, particularly when it is not demonstrably cutting into existing markets or markets that are reasonably likely to be developed.

Of course, copyright owners may be concerned that a revenue-based model will not be adequately compensatory. This would be a grave concern if works were compulsorily licensed. However, a levy system coupled with the opt-out system we proposed in Part III above enables an exit for dissatisfied copyright owners. They can forego their share of the levy, use the opt-out mechanism described in Part III, and then seek a better deal from AI providers, or simply withhold their work. This exit opportunity should also temper AI providers' demands for reduced compensation obligations. Legislators should instead aim for an allocation that broadly satisfies copyright owners, so they are not tempted to opt out and seek higher payments via voluntary licensing agreements.

The interaction between levies and licensing agreements will also be an important topic for calibration of compensation levels. Ed Newton-Rex has announced the development of a certification mark for models using fully licensed content recently.[148] An AI provider that has fully licensed the content it uses should not be required to pay into a levy fund. Similarly, AI providers that have licensed some significant percentage of the works they use should be able to discount their levy obligations commensurately. This would avoid undue compensation for content owners who had already received licensing revenues.

Given the complexities just mentioned, administration of a levy may require a great deal of record-keeping. However, this accounting

[148] Fairly Trained Launches Certification for Generative AI Models That Respect Creators' Rights, Fairly Trained (Jan. 17, 2024), https://www.fairlytrained.org/blog/fairly-trained-launches-certification-for-generative-ai-models-that-respect-creators-rights [https://perma.cc/RP8J-AGLW] (noting that a Licensed Model Certification can be "awarded to any generative AI model that doesn't use any copyrighted work without a license").

for the use of and payment for works may create many spillover benefits. Many leading scholars of AI, like Abeba Birhane and Deborah Raji, have argued that more transparency and accountability is needed with respect to data sets used by AI firms.[149] Their concerns are driven in part by the offensive content in so many databases, but they are also related to demands for fair compensation. Requiring fuller disclosure of works used would be a first step toward achieving the transparency in data sets necessary for several social ends.[150] For example, there is grave concern that certain datasets may not be representative of the populations AI is destined to serve.[151] In response to such problems, transparency better enables public scrutiny of the training data at the core of AI. Hence, the EU Artificial Intelligence Act requires AI providers to publicize sufficiently detailed summaries of content used for training their models.[152]

While other positive externalities are likely consequences of the compensation framework we recommend, we save for later work a full

[149] See generally Inioluwa Deborah Raji et al., Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing (Jan. 3, 2020) (unpublished manuscript), https://arxiv.org/pdf/2001.00973 [https://perma.cc/QNL9-GXTH] (proposing an auditing framework for AI systems); see also Abeba Birhane, Vinay Prabhu, Sang Han & Vishnu Naresh Boddeti, On Hate Scaling Laws for Data-Swamps 15–17 (June 28, 2023) (unpublished manuscript), https://arxiv.org/pdf/2306.13141 [https://perma.cc/P8TE-DFG6] (recommending rigorous audits of large datasets).

[150] Haochen Sun, The Right to Know Social Media Algorithms, 18 Harv. L. & Pol'y Rev. 1, 41-43 (2024) (discussing social values of algorithmic transparency that generalize beyond the social media context).

[151] Id. at 32 ("Given the role of data and data-based inferences in generating discriminatory outcomes, and the black box nature of algorithm design, mandating transparency for input data, rather than algorithms themselves, would likely be more effective in addressing algorithm-generated discrimination and holding platforms accountable.").

[152] European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), para. (107) ("In order to increase transparency of the data that is used in the pre-training and training of general-purpose AI models, including text and data protected by copyright law, it is adequate that providers of such models draw up and make publicly available a sufficiently detailed summary of the content used for training the general-purpose model.").

articulation of them. We have already offered multiple rationales for and modes of valuation of copyrighted works used by AI providers. These methods of valuation are diverse, and may lead to some conflict among stakeholders. But the mere fact that many modes of valuing training data for AI are possible is not an argument for the impossibility of the project. Indeed, the opposite is the case: there are many ways forward. The key now is to begin a vigorous social and political debate on how to value training data, and to expeditiously come to a resolution which respects both the value of AI and the extraordinarily hard work and creativity necessary to create the past and future works on which AI has and will depend.

V. RESPONSE TO OBJECTIONS

Our proposal to couple an opt-out mechanism and a levy may strike some commentators as too favorable to copyright owners or too costly to AI companies. Some have claimed that efforts to compensate authors for use of their work will stop progress in AI.[153] However, it is inconceivable that a modest annual levy would seriously dent the budget of the massive firms behind many of today's leading advances in AI.[154] Smaller levies could be arranged for smaller providers, too. Voluntary licensing is also an "off-ramp" from the levy we propose. OpenAI has already struck licensing deals with leading content providers. [155] Ensuring some level of compensation for creatives will not "break" AI

[153] James Vincent, The Scary Truth About AI Copyright Is Nobody Knows What Will Happen Next, The Verge (Nov. 15, 2022, 10:00 AM), https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data [https://perma.cc/S3UU-9DZW] (describing and contesting such claims).

[154] A levy could also help incentivize the creation of more works for use in AI training, thereby advancing AI in important ways.

[155] Partnership with Axel Springer to Deepen Beneficial Use of AI in Journalism, OpenAI Blog (Dec. 13, 2023), https://openai.com/blog/axel-springer-partnership [https://perma.cc/K5HF-TVHZ]; Madhumita Murgia, The Financial Times and OpenAI strike content licensing deal, Fin. Times, Apr. 29, 2024.

research, just as many online policy reforms will not "break" the Internet.[156]

Another worry is that a critical mass of copyright owners may withhold their work in order to demand higher payments than they would receive as distributions from a levy. If they do so, their withdrawal may seriously impede the further development of AI. There are several responses to such a concern. While legal scholarship commenting on the interpretation of existing copyrighted works has been dominated by analysis of an incentives versus access trade-off, the development of future legislation can, and should, be guided by a more nuanced and inclusive set of policy concerns, including industrial policy.[157] Much depends here on the relative proportion of opt-outs in relation to works as a whole, the importance of such missing works to advances in training generative AI, and the social value of AI in general.

To address the last issue first: While some assume that the development of AI is an unalloyed good, there are numerous indications that the unregulated advance of particular forms of it, including many forms of generative AI, poses threats to privacy and the public sphere.[158]

---

[156] For an example of this "broken internet" complaint, and its refutation, see Danielle Keats Citron & Benjamin Wittes, The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity, 86 Fordham L. Rev. 401, 410 (2017).

[157] This legislative re-orientation would mirror a similar shift in the interpretation of existing antitrust law by regulatory authorities, which has become much more methodologically pluralist over the past decade. Frank Pasquale & Michael L. Cederblom, The New Antitrust: Realizing the Promise of Law and Political Economy, 33 U.S.C. Interdisc. L.J. (forthcoming 2024) (manuscript at 4–5) ("The New Antitrust engages with a wider range of social science expertise [than traditional antitrust] to better inform policy decisions . . . and supplements economic analyses with additional fields of expertise to gain a more holistic view of [the field]."). For a discussion of the incentives versus access trade-off in copyright, see Victor, supra note 146, at 930–35.

[158] See generally Grant Fergusson et al., Elec. Pol'y Info. Ctr., Generating Harms: Generative AI's Impact & Paths Forward (May 2023), https://epic.org/documents/generating-harms-generative-ais-impact-paths-forward/ [https://perma.cc/5HJC-KULQ] (discussing categories of harm posed by AI tools); see generally Johanna Okerlund et al., What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them 62–85 (Apr. 2022), https://stpp.fordschool.umich.edu/research-projects/whats-in-the-chatterbox [https://perma.cc/4KH6-CLQM] (discussing the social and labor costs of LLM adoption); Daniel J. Solove, Artificial Intelligence and Privacy, 77 Fla. L. Rev.

A more controlled, orderly, and restricted transfer of works to AI models may help alleviate such concerns, if only by enhancing the transparency of model construction. Reducing the pace of AI innovation is not an obvious harm; indeed, numerous leaders in the field have signed a letter urging a pause in AI development until more robust systems of regulation could be developed.[159]

Moreover, if a share of the revenues of firms selling generative AI were reserved for compensation for the authors and others whose works it is using, the reduced profitability of the industry could slow down a juggernaut of irresponsible AI applications, such as voice cloning, deepfakes, and the digitization of functions of widely condemned paper mills.[160] In addition, scholars have documented extraordinary environmental harms from AI.[161] Some have even

(forthcoming Jan. 2025) (manuscript at 34), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4713111 [https://perma.cc/D6E7-JBGH] ("The power of AI to make inferences renders many provisions and goals of current privacy law moot.").

[159] Future of Life Inst., Pause Giant AI Experiments: An Open Letter (Mar. 22, 2023), https://futureoflife.org/open-letter/pause-giant-ai-experiments/ [https://perma.cc/7Q9V-FXTN] (calling on "all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4," and observing that "AI developers must work with policymakers to dramatically accelerate development of robust AI governance systems").

[160] On the general case for intellectual property protection as a method of slowing the production and dissemination of social "bads" (to be contrasted with "goods"), see Christopher A. Cotropia & James Gibson, The Upside of Intellectual Property's Downside, 57 U.C.L.A. L. Rev. 921, 921 ("[T]he traditional downside of intellectual property [is] reduced production and impeded innovation. This Article turns the traditional discussion on its head and shows that intellectual property's putative costs can actually be benefits."). On the problem of AI providers acting as paper mills, see Noëlle Gaumann & Michael Veale, AI Providers as Criminal Essay Mills? Large Language Models Meet Contract Cheating Law 7 (Univ. Coll. London Fac. of Laws 2023), https://osf.io/preprints/socarxiv/cpbfd [https://perma.cc/2XU8-W9XT].

[161] Steven Gonzalez Monserrate, The Staggering Ecological Impacts of Computation and the Cloud, The MIT Press Reader (Feb. 14, 2022) https://thereader.mitpress.mit.edu/the-staggering-ecological-impacts-of-computation-and-the-cloud/ [https://perma.cc/W4XL-EQ8K]; Steven Gonzalez Monserrate, The Cloud Is Material: On the Environmental Impacts of Computation and Data Storage, MIT Case Studies in Social and Ethical Responsibilities of Computing 6, 14–16 (Jan. 27, 2022); Kate Crawford, Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence 26–33 (2021).

suggested that the proverbial game is not worth the candle.[162] All of these factors must be weighed against the admittedly great value of AI, as the rights and interests of copyright owners are balanced against those of AI developers and vendors.

The second concern—the relationship between opt-out rights and the available corpus for training—also raises complex questions about the consequences of regulation. In some opt-out regimes governing data, very few persons take advantage of their opt-out rights.[163] The same could occur with respect to copyrighted works and AI. Many copyright owners would likely lack the resources and inclination to withdraw their works from relevant corpora and try to negotiate a better deal with a massive firm like OpenAI.

Creatives may also have moral or other non-monetary objections to the use of their work by certain firms. Yet this should not be interpreted as a rejection of AI *tout court*. Rather, the holdouts may only be seeking to give a commercial advantage to entities more aligned with their own moral commitments, or they may wish to help small competitors of today's AI behemoths. In many cases, this would be an entirely commendable rationale for exercising opt-out rights.

The strength of the first concern raised, regarding the relative proportion of creatives who would opt out, and their decisions' implications for progress in AI, depends in part on still-developing research on the relationship between works' availability and model refinement. A group of experts in the field have developed "a data- and compute-efficient training recipe that requires as little as 3% of the LAION data (i.e., roughly 70 million examples) needed to train existing

[162] Jonathan Crary, Scorched Earth: Beyond the Digital Age to a Post-Capitalist World 8–11 (2022); Dan McQuillan, Resisting AI: An Anti-fascist Approach to Artificial Intelligence 1 (2022).

[163] See, e.g., Lauren E. Willis, Why Not Privacy by Default?, 29 Berkeley Tech. L. J. 61, 97 (2014) ("Although consumers generally do not like banks sharing their information with affiliates or third parties, almost no one opts out.").

SD2 [Stable Diffusion 2] models, but obtains the same quality."[164] They conclude that these "results indicate that we have a sufficient number of CC [Creative-Commons-licensed] images (also roughly 70 million) for training high-quality models." [165] While trade-offs between data availability and model quality may persist in all these areas, the *summum bonum* of copyright policy is not the maximum advance of AI. Creatives' interests must also be taken into account. [166]

Unexpected interactions between works' availability and advances in computer science may also occur. Restrictions on free access to extant copyrighted works may lead to advances in computational efficiency designed to do more with less. If neural network-based approaches premised on the "unreasonable effectiveness of data" [167] experience reduced quality because of copyright restrictions, this may simply accelerate what some commentators deem a long overdue shift toward alternative approaches, including more explainable, symbolic, or neurosymbolic AI.[168] Even if this fails to occur (or leads to a research dead end), the potential reallocation of computing talent in the wake of copyright-induced challenges to present industry leaders is by no means necessarily problematic. Those now working on perfecting AI-generated music, movies, and novels may turn their considerable talents to advancing computation in less copyright-intensive areas, such as

[164] Aaron Gokaslan et al., CommonCanvas: An Open Diffusion Model Trained with Creative-Commons Images 1 (Oct. 25, 2023) (unpublished manuscript), https://arxiv.org/pdf/2310.16825.pdf [https://perma.cc/WW52-M65H].
[165] Id. at 2.
[166] See Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 531 (2023) (describing the "goal of copyright" as promotion of "the progress of science and the arts, *without diminishing the incentive to create*.") (emphasis added).
[167] See generally Alon Halevy, Peter Norvig & Fernando Pereira, The Unreasonable Effectiveness of Data, IEEE Intell. Sys., Mar.-Apr. 2009, at 8 (describing the use of data in natural language learning).
[168] For arguments for alternative approaches to AI development, see Gary Marcus, Deep Learning Alone Isn't Getting us to Human-Like AI, Noema Mag. (Aug. 11, 2022), https://www.noemamag.com/deep-learning-alone-isnt-getting-us-to-human-like-ai/ [https://perma.cc/D74N-XUEN]; Gary Marcus & Ernest Davis, Rebooting AI: Building Artificial Intelligence We Can Trust 41–44 (2021).

medicine, agriculture, and logistics. It is far from clear that such a shift in investment would harm society.

In short, it is exceedingly difficult even for those within the AI field to forecast the medium- and long-term effects of the changes in relative costs of data that our proposal would likely bring. Armchair consequentialism can neither invalidate nor prove the value of our proposal. Rather, uncertainty here commends a principle-centered, rather than results-centered approach, while policymakers also continually re-evaluate the effects of legislative adjustment of rights and interests. The principles of consent and compensation are our lodestar and are designed to protect the legitimate interests of copyright owners while not unreasonably prejudicing the advance of AI innovation.

V. CONCLUSION

Faced with untrammeled expropriation of their works by AI providers, creatives have demanded consent, credit, and compensation.[169] In terms of consent, they want the ability to refuse the inclusion of their works in databases used by AI providers. In terms of credit, they want to overcome the pervasive and deeply troubling trade secrecy now so characteristic of AI development in order to discover whether their works were used for training models and generating content. Compensation has been less well-specified, but means some fair share in the revenues created by an AI market potentially valued in the trillions of dollars. This essay has directly addressed concerns about consent and compensation, while indirectly promoting proper attribution of credit by advancing a mechanism designed to expose and remedy infringement.

By proposing a scheme for addressing creatives' concerns, this essay has made at least three contributions. First, we have made the

[169] See supra note 11 and accompanying text.

case for coupling control and compensation mechanisms, whereas copyright law in the past has tended to develop one at the expense of the other. Second, we have developed a suite of rationales for legislative change that focus on the avoidance of unjust enrichment. Third, we have proposed rationales and levels of compensation that may serve as benchmarks for further development by policymakers and negotiation by stakeholders.

In short, it is time for a New Deal with respect to copyright and AI. Numerous lawsuits against AI providers are forcing policymakers around the world to rethink an increasingly broken social contract between technologists and creatives. The time is right for a legislative solution along the lines we have proposed.