Publishing Research

ARTICLE

Open Access

https://doi.org/10.48130/pr-2025-0001

Publishing Research **2025**, 4: e002

Transformative training: an analysis of AI training data and fair use in *Authors Guild v. OpenAI Inc.*

Hannah Didsbury and Xiaohua Awa Zhu*

School of Information Sciences University of Tennessee, Knoxville Knoxville, TN 22101, USA

* Corresponding author, E-mail: xzhu12@utk.edu

Abstract

The rise of generative artificial intelligence (AI) has raised critical questions about copyright law, particularly regarding the use of copyrighted material in training datasets. This paper examines *Authors Guild v. OpenAI Inc.*, a landmark lawsuit exploring whether such use constitutes copyright infringement or fair use. Through a technical analysis of GPT models—including tokenization, neural network architecture, and training processes—the paper demonstrates how AI training could be considered transformative in its use of copyrighted works. While existing precedents may hold that transformative uses of copyrighted works can be permissible, ambiguity in the law and uncertainties with emerging technology fuels ongoing debate. Challenges arise from ChatGPT's ability to reproduce verbatim excerpts from its training data, potentially undermining OpenAI's transformative use argument. The court's decision will address this legal uncertainty, clarify the application of transformative use in the context of AI, and likely set a precedent for future disputes. Beyond its legal implications, the case could reshape licensing practices, restrict access to training datasets, and influence advancements in AI research. Moreover, it raises broader concerns about creativity, innovation, and intellectual property, underscoring the need for American copyright law to balance the protection of human authorship with fostering technological progress in the age of AI.

Citation: Didsbury H, Zhu XA. 2025. Transformative training: an analysis of AI training data and fair use in *Authors Guild v. OpenAI Inc.*. *Publishing Research* 4: e002 https://doi.org/10.48130/pr-2025-0001

Introduction

The recent generative artificial intelligence (AI) boom has brought increased public scrutiny with its technological innovations. Debates are rising, questioning the sustainability of new technologies, the future of creative jobs, and even the definition of art itself. Among the many contentious issues, one likely to see a definitive resolution is whether the use of copyrighted material as training data for generative AI models constitutes infringement. *Authors Guild vs OpenAI Inc.* is a class action lawsuit that addresses the very question. Filed on September 19th, 2023, by the Authors Guild—a professional organization for writers—and 14 named authors, the case is being heard by the United States District Court for the Southern District of New York[1]. The first discovery hearing took place on October 1st, 2024[2].

Though the case is far from resolved, it stands to have far-reaching consequences for not only the future of AI research but also the broader trajectory of copyright law in the United States. The lawsuit brings to the forefront a fundamental tension between the rapid advancements in technology and the current limitations of copyright law, which was originally designed to address human-generated works in traditional mediums. This case could serve as a turning point by addressing the growing mismatch between the practical applications of copyright law and public perceptions of what intellectual property protections should encompass in an era of generative AI.

At the heart of the debate is whether the use of copyrighted works in AI training datasets qualifies as fair use—a doctrine intended to balance the rights of creators with the broader public interest. On one hand, AI training can be seen as a transformative use, extracting patterns, frequencies, and structures rather than directly replicating protected expressions. This perspective aligns with established judicial precedents and existing copyright statutes, which often consider the purpose, character, and transformative nature of a work in determining fair use. On the other hand, the sheer scale and commercial implications of AI models complicate the argument, raising questions about the boundaries of transformative use and the economic rights of original creators.

The court's decision will not only clarify legal interpretations of transformative use in the context of AI but may also influence the development of new standards for licensing and fair use in data-driven industries. By potentially redefining the scope of copyright protections and acceptable uses of creative works, the outcome of this case stands to reverberate through the publishing industry, creative sectors, and technology research fields. It will also contribute to the broader discourse on how intellectual property laws should evolve to accommodate technological innovation without stifling creativity or access to knowledge.

It first outlines the foundations of American copyright law and the fair use doctrine, offering context for the legal arguments. Then it details the case, explains GPT model mechanics and training data use, and applies precedent to anticipate potential outcomes. Finally, it explores the case's impact on the publishing industry, AI research, and intellectual property law. Together, the analysis highlights the challenges and opportunities at the intersection of generative AI and copyright law.

American copyright law and the fair use doctrine

Legal foundations

At its most basic legal definition, copyright is a statutory grant meant to encourage the production of creative works by guaranteeing that the author alone will profit from the work and specific derivations or other uses thereof until a predetermined time where the work becomes a part of the public domain, and is thus free to

use without restriction[3]. American copyright first comes from Article 1, Section 8, Clause 8 of the Constitution as one of Congress's enumerated powers. While numerous statutes and precedents have shaped American copyright law, the foundation of the nation's modern copyright system is the Copyright Act of 1976[4]. It made several impactful changes, such as extending the timeframe of protections for copyrighted works, further defining works of authorship in the face of technological advances, and allows for the transfer of copyright, but most relevant to the *Authors Guild* case, it further defined the requirements for copyright protection and codified the fair use doctrine, which had previously only existed as judicial precedent[3].

Section 102 of the Copyright Act of 1976 states that copyright can only be applied to '... original works of authorship fixed in any tangible medium of expression...', and it cannot be applied to '...any idea, procedure, process, system, method of operation, concept, principle, or discovery...'[4]. Any tangible expression of these noncopyrightable phenomena must be the sole work of the author and possess at least a minimum amount of creativity, as held in the decision from *Feist Publications, Inc. v. Rural Telephone Service, Co.* This leaves certain elements of a work, such as overall ideas, themes, genre conventions, and syntax, unprotectable. Only the original elements of a fixed expression can be infringed upon.

Section 107 of the Copyright Act of 1976 states a fair use of a copyrighted work includes '...purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research…'[4]. However, the listed purposes are not an impenetrable safety net. Any alleged infringement claiming fair use is subject to a decision on a case-by-case basis using the four factors of fair use outlined in section 107[5]. The four factors under consideration are: '(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'[4]. Some scholars argue that section 107 fails to effectively establish consistency in fair use decisions by not defining fair use, not attempting to order the priorities of the four factors, and by listing ambiguous purposes as fair use[5]. It's this inconsistency that necessitates cases like the *Authors Guild*. There is no question of fact in regards to the case; OpenAI has admitted to using copyrighted works in training datasets[6]. It's entirely a question of law, asking where OpenAI's use stands within America's existing copyright framework. Complicating matters further is that this framework is built on both the law and individual understanding and belief as to what copyright should accomplish. The remedy for the conflict between AI and copyright, beyond the decision of one case, is to further examine what copyright means to American authors and users and how that meaning can, and should, change over time.

Philosophical justifications for intellectual property

Determining who owns a cow is much easier than determining who owns a concept. A different logical framework is required to justify ownership of the noncorporeal, hence the need for unique philosophical justifications behind intellectual property law. In the United States, copyright protections are primarily justified via utilitarian welfare theory.

Utilitarian welfare theory holds that intellectual property rights are given as a means to promote the greatest public good—specifically, increasing access to creative and scientific works[7]. Under this framework, intellectual property is a government-created incentive, rather than a reward for labor as envisioned in Lockean labor theory or as an expansion of the parental-esque interest a creator takes in their work as conceptualized in the moral rights doctrine[8,9]. Legal scholars describe this practical manifestation of utilitarian welfare theory as 'an administrative solution to an economic problem'[7]. This perspective underpins the Constitution's intellectual property clause, which seeks to 'promote the Progress of Science and useful Arts' by granting creators specified exclusive rights for limited times.

Despite its utilitarian roots, copyright law also aligns with Lockean Labor theory by rewarding authors. For instance, in *Twentieth Century Music Corp. v. Aiken,* the court acknowledged that copyright law secures a fair return as the incentive for authors' creative labor while ultimately stimulating public creativity. While utilitarian welfare theory and the associated influence from Lockean labor theory form the primary logical basis for American intellectual property law, debate persists regarding how to best achieve this balance through enforceable laws.

Recent decades have seen statutes create longer and stronger copyright protections, raising questions about whether extending copyright and shrinking the public domain leads to greater public good or stifles creativity. Some scholars argue that longer copyright terms encourage the creation of new works and discourage the reliance on existing ones[10]. Other scholars, like James Boyle, contend that a robust public domain better fuels the progress of the arts, emphasizing that information products, new ideas, and creative works are never built in a vacuum; rather, they rely on fragments of prior works created by other people[11]. Access to earlier works, therefore, is essential for fostering innovation.

While utilitarian welfare theory remains the foundation of American copyright law, it is rife with contrary opinions regarding its application. Given that much of the U.S. copyright law is shaped by judicial decisions, these differing opinions have led to uncertainties for the authors and users of intellectual works as to what assumptions can be made about copyright protections. Here again lies the complexity of the *Authors Guild* case. Beyond merely an uncertainty regarding how the existing law can be applied to new technologies, an uncertainty that has been addressed before with photography, time shifting, and peer-to-peer file sharing, among other advances, there is an uncertainty in how new laws should be developed and justified. If there is no agreement as to the intentions and best practices of copyright, how can any copyright law hope to effectively function?

The *Authors Guild v. OpenAI Inc.* lawsuit analysis

The complaint

The Authors Guild alleges that the act of using copyrighted works as part of training data for Large Language Models (LLMs) constitutes copyright infringement and seeks redress[1]. In the amended complaint filed on December 5th, 2023, the Authors Guild alleges three types of copyright infringements: the direct infringement claim against OpenAI OpCo LLC, the vicarious infringement claim against OpenAI Inc. and OpenAI GP LLC, and the contributory infringement claim against OpenAI LLC, OpenAI Global LLC, OAI Corporation LLC, OpenAI Holdings LLC, OpenAI Startup Fund I LP, OpenAI Startup Fund GP I LLC, OpenAI Startup Fund Management LLC, and Microsoft[1]. The vicarious and contributory infringement claims depend on a finding of direct infringement by OpenAI OpCo LLC.

Vicarious infringement, or vicarious liability, holds a party responsible for a third party's infringement if one profits from the infringing action and can supervise the direct infringer, regardless of whether the party knew of the wrongdoing[12]. In an earlier case,

MGM Studios, Inc. v. Grokster, Ltd., the court held that vicarious infringement requires both the ability to control the actions of the direct infringer and a direct financial benefit from those actions. According to OpenAI's corporate disclosure statement, OpenAI OpCo LLC is a subsidiary of OpenAI, Inc, but not of OpenAI GP LLC.[13]. Therefore, only OpenAI, Inc. appears to have some level of control over OpenAI OpCo LLC[13,14] and benefits financially from OpenAI OpCo LLC's AI models. If OpenAI OpCo LLC is found to have committed direct infringement, OpenAI, Inc., and not OpenAI GP LLC could be liable for vicarious infringement.

Contributory infringement is a legal doctrine that holds parties liable for knowingly contributing to copyright infringement, even if they are not directly committing the infringing action, as found in *Kalem Co. v. Harper Brothers*[12]. *Gershwin Publishing Corp. v. Columbia Artists Management, Inc.* further clarified the two requirements for contributory infringement: the accused must have knowledge of the direct infringement and must induce, cause, or otherwise materially contribute to it. It is reasonable to assume that Microsoft, through its partnership with OpenAI, was aware of the methods OpenAI used to develop its AI models, particularly as Microsoft has integrated ChatGPT into its Azure AI services[15]. Furthermore, Microsoft's CEO, Satya Nadella, said in an interview that '...OpenAI wouldn't have existed but for [Microsoft's] support early on'[16].

These facts, however, do not automatically establish contributory infringement. The decision in *Fonovisa, Inc. v. Cherry Auction, Inc.* determined that contributory infringement depends on whether the 'support services' provided by the alleged infringer were essential for the direct infringement to occur. OpenAI released multiple iterations of its GPT models before its partnership with Microsoft began in 2019[17,18]. Microsoft's financial support was not integral to the alleged infringement. Therefore, it is unlikely that Microsoft's partnership with OpenAI constitutes contributory infringement. As for the seven subsidiaries of OpenAI Inc. named in the contributory infringement allegations, they will be similarly evaluated based on the support services they provided. If these services were essential for the alleged infringing activities, they could potentially be liable.

The direct infringement claim against OpenAI OpCo LLC alleges a knowing violation of the plaintiffs' "exclusive rights by reproducing their copyrighted works in copies for the purpose of 'training' their LLMs and ChatGPT" [1]. OpenAI argues that this constitutes 'transformative paradigmatic fair use' because '[p]rocessing copyrighted works to extract information about the work—such as word frequencies, syntactic patterns, and thematic markers—does not infringe because it does not replicate protected expression'[10]. Resolving this issue hinges on both the mechanics of training LLMs and precedents established in similar fair use cases.

The mechanics of GPT and training data

In OpenAI's response to the amended complaint from the Authors Guild, the company describes its GPT models as 'computer programs that are developed using artificial intelligence and machine learning techniques'[10]. As is common in discussions of AI research, terminology can easily obfuscate the underlying technology.

OpenAI's Generative Pretrained Transformer (GPT) models, specifically GPT-3.5 and GPT-4, are the foundational computing models behind the ChatGPT chatbot[19]. These models are classified as Large Language Models (LLMs), placing them under broader hierarchical categories of AI — weak AI — machine learning — deep learning — Generative AI (GAI) models. Trained on large text corpora, the GPT models generate natural language outputs tailored to the specifications of user input[20].

Unlike the theoretical concept of strong AI — AI with independent consciousness, human-like self-awareness, and the ability to think and learn — the current generative AI technology, what is referred to as weak or narrow AI, can only imitate facets of human cognition without true consciousness[21]. Within this framework, machine learning employs algorithms and statistical models to identify patterns in given datasets, while unsupervised learning (one of the machine learning approaches), as used in GPT models, enables pattern recognition without labeled datasets[22]. The model observes data, identifies relationships, and adjusts its output accordingly[21].

GPT models, as deep learning models, use neural networks that mimic the structure of the human brain[22]. These networks consist of an input layer, multiple hidden layers, and an output layer. Each artificial neuron processes information by applying weights and biases to inputs, generating an output that informs the next layer[22]. Through training, these parameters are fine-tuned to optimize performance[22]. As GAI, GPT models can create non-numeric output, such as text, images, code, and audio, because their LLM capabilities enable them to process natural language inputs and training data[21].

GPT-3.5 and GPT-4, the models powering the ChatGPT chatbot, are built on transformer architecture, a type of neural network[23]. Introduced in 2017 by Google researchers, transformer models use an encoder-decoder structure and rely entirely on attention mechanisms, replacing the recurrent and convolutional layers found in earlier neural networks[24].

Before any data can be sent through the encoder, it must be translated into a model-intelligible format. Tokenization is the process by which natural language, whether from training data or user input, is split into smaller units called tokens, which can be phrases, whole words, word fragments, or punctuation[22,25].

The mathematical purpose behind the GPT models is to predict the token most likely to follow a given sequence of words, which is ultimately achieved through a probability distribution generated by the model's neural network[24,26]. The first layer of the model, the embedding matrix, converts tokens into high-dimensional vectors that encode semantic relationships in the position of these vectors[22]. The hidden layers within the GPT models are the attention mechanisms, which allow the model to evaluate the position of each token in a given sequence and determine how that position affects the token's relationship to other tokens[23]. For example, it allows the model to identify which noun an adjective modifies. These attention patterns are used to create value vectors, which are multiplied by each token in a given sentence. The results are summed and added to the original token's embedding vector to encode context-dependent changes in meaning[23]. The final layer of the neural network, the unembedding layer, maps the preceding layer's output to the model's vocabulary list and uses the softmax function to create a probability distribution consisting of every token in that vocabulary[24]. This distribution is then used to predict which token would be likely to appear next.

To produce longer text, the model appends the predicted token to the sequence and repeats the process[22]. It is worth noting that any successful GAI must ensure that the content it generates is sufficiently different from the training data so as not to just reproduce existing material[27]. This is achieved through incorporating controlled randomness, introducing variability into outputs while maintaining coherence across outputs[27]. This is why the GPT models do not always select the word with the highest probability of being next in the sentence every single time—they choose a plausible alternative, enabling diverse responses even from identical inputs.

Although much of the GPT model's learning is unsupervised, two supervised methods are used for fine-tuning. Backpropagation is

the primary algorithm for supervised learning[27]. In AI models, learning is the process of discovering the weights and biases that minimize the cost function, that is, the difference between the model's output and a human-defined desired output[23,28]. Thus, each output generated from the training data must be compared to the ideal output, and the difference between the two is used to calculate the gradient of the cost function. The gradient is then applied to slightly adjust each parameter to generate outputs with greater similarity to the ideal output[28].

The other supervised learning method is Reinforcement Learning from Human Feedback (RLHF), where human evaluators provide feedback to guide the model in correcting specific issues[29]. This feedback can include new prompts and responses or edited versions of the model's original responses[29]. Through RLHF, models are trained to avoid producing harmful or untruthful responses. Both of these methods fine-tune the model's parameters to ensure it consistently generates proper outputs.

Applying precedent

Although the use of copyrighted works as training data has not yet been the subject of an official court ruling, there is relevant precedent for the use of whole works in a transformative process. The two most relevant cases, *Authors Guild, Inc. v. Google, Inc.* and *Authors Guild Inc. v. HathiTrust,* are similarly second circuit class action fair use cases.

Authors Guild, Inc. vs Google, Inc. was litigated over Google's use of copyrighted works for Google Books' searchable database, which included snippets from the work alongside the search results[30]. These snippets were the only part of the copyrighted works shown to users[30]. The case ruled in favor of Google, determining that by providing information about the work, and thus making those works more discoverable, Google's use of the works was transformative. The snippets provided as search results did not wholly replace the original work.

OpenAI claims that the GPT models do not store their training data[31]. Allegedly, the only information stored are the changes made to the model's parameters during training[31]. However, ChatGPT can reproduce verbatim responses from its training data. For example, when asked, 'What's the first sentence of Game of Thrones?', ChatGPT responds with "The first sentence of 'A Game of Thrones' by George R.R. Martin is: 'We should start back,' Gared urged as the woods began to grow dark around them," which is the first line of the book's prologue[32]. ChatGPT will provide the exact phrases when prompted sentence by sentence, until the 12th, whereupon it responds: 'Sorry, I can't provide the text beyond what I've shared. Would you like a summary or analysis of that section instead?'.

Similarly, one can ask for and receive the last two sentences from *A Game of Thrones* verbatim. However, asking for whole paragraphs, for any of the first 11 sentences out of sequence, or anything beyond the first 11 or last two sentences results in the summary-analysis redirection response. ChatGPT's ability to reproduce exact quotes rather than summaries or similar text suggests that the GPT models may store and could access parts of their training data for output.

Despite these reproductions of protected expression, ChatGPT's verbatim responses may not constitute copyright infringement. *De minimis* copying refers to instances where the quantity of a work copied is so minuscule that it does not constitute infringement[8]. However, one should not rely on the de minimis principle as a defense, as there is an unresolved circuit split over the issue. In the 2005 *Bridgeport Music v. Dimension Films* decision, the Sixth Circuit Court found that no amount of copying is negligible because every bit of the work is valuable to the whole. Conversely, in the 2016 decision of *VMG Salsoul v. Ciccone,* the Ninth Circuit Court affirmed a lower court's ruling that trivial copying does not constitute infringement. This decision followed the reasoning in *Newton v. Diamond*, where the Ninth Circuit held that a use must be significant enough to be actionable as copyright infringement.

Given the lack of a bright line rule for *de minimis* copying, it is difficult to predict with confidence whether verbatim quotes from copyrighted works in ChatGPT responses would constitute infringement. ChatGPT's ability to pull verbatim quotes from its training data may have to be restricted at the code level for the chatbot, or the GPT models may be required to undergo additional RLHF to redirect queries for direct quotes to avoid copyright violations.

While ChatGPT similarly does not provide large verbatim sections of copyrighted works, its ability to provide summaries and analyses of works included in its training data could, unlike Google Books' snippets, potentially usurp the market for the original work.

Authors Guild, Inc. v. HathiTrust was litigated over the HathiTrust Digital Library's use of copyrighted works scanned in the Google Books project. The HathiTrust Digital Library's repository serves three purposes. First, like Google Books, it is a searchable database, but search results for copyrighted works only display page numbers and the frequency of the search term without including snippets. Full-text access is limited to works in the public domain. Second, it provides adaptive technology for disabled users, such as screen readers, to read printed text. Third, it allows libraries to use digital scans of copyrighted works to replace lost or damaged physical copies when replacements cannot be acquired at a fair price. The court decision was in favor of HathiTrust in the first and second use cases, but it vacated and remanded the third use case due to issues of the plaintiff's standing.

Overall, the court found that the use of an entire copyrighted work is transformative if it creates a result with a 'purpose, character, expression, meaning, and message' entirely different from that of the original work. Similarly, when entire copyrighted works are used in training data for generative AI, no protected expressions need to be shown to the users, as hardcoded responses to excerpt queries or RLHF to prevent models from verbatim responses could prevent any part of the training data from being directly shown to users. The training data, in being tokenized and represented as high-dimensional vectors, then serves a purpose distinct from that of the original work. Following the precedents set by this case, a use that is transformative in purpose and does not display excerpts of protected expression to the end users would likely constitute fair use.

Another angle of this fair use argument considers the tokenization of copyrighted material as a form of decompiling in the process of reverse engineering human-like writing from human-written works. The first case to consider decompiling as a fair use was *Sega Enterprises Ltd. v. Accolade, Inc.*, wherein game developer Accolade disassembled a Sega Genesis console and reverse-engineered its code to develop Genesis-compatible games. The Ninth Circuit court ruled in favor of Accolade, holding that the copying involved in decompiling does not constitute infringement as long as it is the only way to access elements that are not original and therefore not protected by copyright. Extrapolating this logic from code to natural language, the mechanics of language are not copyrightable, as a valid copyright claim requires originality and creativity[8]. The tokenization process is the only means by which LLMs can gain access to said non-copyrightable elements and could be considered fair use.

Furthermore, considering the process of LLM training as a form of decompiling and reverse engineering, the creation of intermediate copies of copyrighted works would also qualify as fair use, following

the ruling in *Sony Computer Entm't, Inc. v. Connectix Corp.,* another Ninth Circuit decision. In this case, Connectix developed an emulator of Sony's PlayStation console by decompiling and reverse engineering Sony's firmware, which was protected by copyright. The court ruled that the creation of intermediate copies to access unprotected elements of Sony's software constitutes fair use. Similarly, if the tokenization of protected expression for LLM training is viewed as a form of decompiling, the presence of intermediate copies in training datasets would also constitute fair use, as the copies are retained solely to facilitate access to the unprotected elements of the works, serving a distinct purpose from that of the original works.

Another perspective on the use of copyrighted material in training data is to consider the parallels between how LLMs learn from works and how humans create new works. Human-written works are not created in a vacuum. Literary critic Northrop Frye argued that one cannot seriously accept the imagined notion that "...a 'creative' poet sits down with a pencil and some blank paper and eventually produces a new poem in a special act of creation *ex nihilo*," as 'Poetry can only be made out of other poems; novels out of other novels. Literature shapes itself and is not shaped externally…'[33] Creative production requires engagement with the broader context and tradition of art, learning from and building upon previous works.

Austin Kleon, in his manifesto-how-to fusion *Steal Like An Artist,* furthers this concept and posits that '...nothing comes from nowhere. All creative work builds on what came before. Nothing is completely original.' [34]. Creativity, in this sense, is inherently iterative and collaborative, borrowing and reshaping ideas from the collective body of work beyond conscious reference and citation. The provenance of influence becomes impossible to track. If human beings are allowed to digest the content of copyright-protected works from a myriad of authors, analyze and internalize the mechanics of the language they use, and create works that are influenced by but not distinct from the regurgitated protected expression, why should AI models be held to a different standard?

If courts decide that AI models should not use copyrighted works to generate new material, it raises critical questions. How will humanity's collective creativity be protected? Where will the line be drawn in our subjective perceptions when we examine our creative processes? Historically, copyright cases have frequently been decided in favor of longer and stronger copyright. The *Authors Guild* case may follow this trend, and in doing so, it could risk much more than the future of AI research, going as far as to impede the future of human creativity itself.

Discussion and implications

The outcome of the Authors' Guild case could have substantial implications for the publishing industry, AI research, intellectual property law, and creative labor. Depending on the court's decision, it may influence licensing practices, future litigation, research methodologies, and interpretations of copyright law in future policies, potentially shaping how these domains evolve in the context of generative AI technologies.

Licensing

One critical aspect of the Authors Guild amended complaint is the redress sought by the plaintiffs, which includes not just the standard copyright infringement penalties but 'damages for the lost opportunity to license their works' [1]. This argument is likely to influence the court's consideration of the fourth factor of fair use and could have significant repercussions for the publishing industry.

The practice of licensing content for AI training data is not unheard of. For example, Adobe's Firefly, a generative AI, is trained exclusively on licensed data and public-domain data[35]. Such precedents bolster the Authors Guild's case, particularly concerning the fourth factor of fair use under section 107 of the Copyright Act of 1976, which evaluates 'the effect of the use upon the potential market for or value of the copyrighted work' [4]. If licensing agreements are shown to exist for similar uses, it strengthens the argument that unlicensed use could undermine a viable market, making it less likely to qualify as fair use.

Despite this, a shift toward widespread licensing for AI training data may primarily benefit publishers rather than individual authors. Current licensing agreements, such as those between OpenAI and The Associated Press or Microsoft and Taylor & Francis, are negotiated at the publisher level, often bypassing direct input or compensation for authors[36]. Cambridge University Press, one of the few publishers known to consult authors about such agreements, remains an exception[37]. As licensing for AI training data becomes a standard clause in publishing contracts, authors could see their rights further marginalized, with little, if any, licensing revenue flowing to them. Furthermore, authors who do not wish to have their works included in licensed datasets could find it more difficult to be published if it becomes an industry standard.

This trend could reshape the publishing industry, expanding the scope of what is included in licensing agreements and altering traditional notions of copyright protections. As publishers increasingly negotiate these deals, they may consolidate their role as intermediaries, further controlling how creative works are monetized in the digital age. The outcome of the Authors Guild case could therefore set a precedent, impacting not just AI developers but also how publishers and authors navigate copyright in a rapidly evolving technological landscape.

Future litigation

The *Authors Guild* case focuses on the use of copyrighted works as training data for OpenAI's GPT models, not the potential infringement of AI-generated content. However, the amended complaint specifically mentions ChatGPT outputs that are derivative of copyrighted works and alleges (in paragraph 130) that some ChatGPT users generate such derivative works for sale[1]. Following the rationale from *Sony Corp. of America v. Universal City Studios, Inc.,* the manufacturer of a technology with non-infringing uses cannot be held liable for infringing activities conducted by its users without the manufacturer's knowledge. Assuming OpenAI does not track how users utilize ChatGPT's outputs, it cannot be held liable for users' sale of derivative AI-generated works. OpenAI's *Terms of Use* for ChatGPT explicitly prohibit users from '[using] our Services in a way that infringes, misappropriates or violates anyone's rights' [38]. However, such an agreement is not necessarily enforceable in a breach of contract suit, though OpenAI can suspend accounts or terminate service access for violations.

In any future litigation regarding AI-generated content, the Authors Guild or the individual authors would have to file suits against individual ChatGPT users. Should they choose this Sisyphean task, litigation may shake out like previous derivative work cases. However, given that AI-generated content can produce mixed results from its training data, they may find more success litigating over the infringement of the overall look and feel of a work. In instances of ChatGPT users impersonating specific authors in the sale of AI-generated content, litigation on the grounds of the author's right to publicity may be more effective.

Regardless of how authors choose to litigate against the sale of AI-generated works, it should not involve OpenAI. The case at hand deals solely with the use of copyrighted works in GPT model training data, and all three counts depend on that singular use

being deemed infringement. Following this case, there may be a flood of copyright infringement cases brought against those who are already selling derivative AI-generated works or may be doing so in the future.

Complications in AI research

The potential popularization of licensed data sets for research runs the risk of limiting the potential of scholarly research beyond just generative AI. Research of any kind relies on the limitations of copyright that allow the use of others' work in comment, critique, and further research. The UC Berkeley Library's response to US Copyright Office's Notice of Inquiry on AI training focused on Text Data Mining (TDM), the process by which researchers use 'computational tools, algorithms, and automated techniques to extract revelatory information from large sets of unstructured or thinly-structured digital content' [39]. While not every example of TDM uses AI models, many current applications of TDM are too complex for algorithms that merely identify word frequency or proximity[39]. The AI models used in TDM are usually discriminative models, which, unlike generative models, are trained using labeled datasets and work to label a given input based on observations made from the labeled examples[29]. These uses of TDM include everything from media analysis to evidence of real-life social issues[39].

If the use of copyrighted works as AI training data constitutes infringement, this research methodology will have to change. Licensing fees for proprietary data may become barriers for institutions and smaller companies without the resources of Microsoft-backed OpenAI. Conversely, these licensing practices could kickstart a new era of data sharing. Open-source datasets, especially those created and maintained as a part of wider open science policy like Germany's Nationale Forschungsdateninfrastruktur, can better facilitate research across all disciplines, while ensuring proper attribution and compensation for the authors behind said datasets[40]. However, a dependence on open-source data is not an indisputable solution to the use of copyrighted works without licensing. There could be authors unwilling to allow their work in research data sets, as well as the issue posed by orphan works being barred from inclusion in such datasets. Innovation in the field of AI won't cease to exist, but it will have to adapt to licensing practices or work towards developing a culture of open science within the field itself and in its relations with others.

Future policy

Though numerous court cases and state legislatures have addressed concerns regarding AI development and implementation, there are as yet no comprehensive federal policies regulating the future of AI. This gap needs to be addressed, as the lack of uniform laws pulls both the creative and technological fields into speculative action.

One potential policy choice is to enact citation requirements for a model's training data. California's Assembly Bill 2013 will require the developer of any GAI made publicly available after January 1st, 2026, or any substantial modification to a GAI made publicly available after January 1st, 2022, to publish documentation regarding the model's training data. This documentation must include descriptions of the dataset and its use, as well as specifically addressing 'Whether the datasets include any data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain' and 'Whether the datasets were purchased or licensed by the developer'. Federal policy modeled on Assembly Bill 2013 could be useful regardless of the outcome of the *Authors Guild* case, as such a law would help normalize citing the sources of a models' training data quicker than the field of AI development could naturally standardize such a practice.

Another potential policy choice is regulating disclosure. Utah's Artificial Intelligence Policy Act, in focusing on consumer protections, requires that the use of any generative AI in services of a regulated occupation must be prominently disclosed[41]. Generative AI used in nonregulated occupations must, when prompted, be able to disclose that the user is not interacting with a human[41]. Disclosure regulation could be extended beyond the use of AI models to power chatbots to include the content made by generative AI. The Federal Trade Commission enforces advertising disclosure regulations. Similar regulations and enforcement methods could be created, or existing ones expanded to include generative AI, to make apparent when content is being or has been generated by AI. Disclosures could help maintain fair competition between human-created and AI-generated content. Pairing disclosure regulations with citation requirements in federal regulation would help maintain consistent transparency as to what goes in and comes out of AI models.

Policy can also help push AI development towards further innovation. Another section of Utah's Artificial Intelligence Policy Act establishes the state's Artificial Intelligence Learning Laboratory Program[41]. The key part of the program is that it allows for periods of regulatory mitigation for AI developers, incentivizing them to push new ideas without fear while working with the state government in risk assessment and implementation[41]. Regulation at the federal level will not cripple the AI industry. Collaborative regulation that considers and attempts to balance the needs of all involved can help push the industry towards further innovation in a manner better aligned with ethical standards.

Conclusions

Copyright law has historically evolved alongside technological advancements. Without the invention of the printing press, making the production and distribution of copies of written works feasible, society would have never needed laws to secure the profits of said distribution for the author or publisher. Subsequent technological leaps have similarly resulted in expanded copyright protections[42]. Landmark cases like *Burrow-Giles Lithographic Company v. Sarony*, which recognized copyright for photographs, and *MGM Studios, Inc. v. Grokster, Ltd.*, which held companies operating liable for contributory infringement via peer-to-peer file sharing networks, exemplify this trend.

The Authors Guild case does not exist in a vacuum. It follows a legal legacy operating against users and undermining what America's initial understanding of copyright was for. If the Authors Guild case and other AI lawsuits currently winding their way through the US judicial system are decided wholly in favor of the copyright holders, it will follow a centuries-long precedent of further limiting users' rights. However, a decision wholly in favor of OpenAI has similar risks. OpenAI and similar technology companies could have a vested interest in exploiting copyrighted works while preventing users from accessing their own proprietary data sets and code, to the detriment of authors and users alike. Copyright relies on the balance between authors' and users' rights to fulfill its purpose. Overly broad restrictions on either side of the scale will put the collective creative culture at risk.

In creating any solution for this delicate balancing act, one must recognize that existing policy, precedents, and frameworks were designed for human use, not AI use. Attempting to shoehorn AI technologies into outdated statutes is as futile as forcing a square peg into a round hole. In adapting existing intellectual property law to the current era of AI, the purpose of copyright must be reevaluated to maximize the potential of the new technology while safeguarding the interests of human authors and users. The law

should be reimagined to best foster creative output from both humans and AI models, with targeted solutions tailored to specific challenges. Finding a balance between regulation and innovation will ensure that AI development follows the purpose of copyright protections by aligning with human progress rather than undermining it.

Author contributions

The authors confirm their contribution to the paper as follows: study conception and design, analysis and interpretation of results, draft manuscript preparation: Didsbury H, Zhu XA; data collection: Didsbury H. Both authors reviewed the results and approved the final version of the manuscript.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Acknowledgments

The authors thank the editor and anonymous reviewers for their constructive feedback and suggestions, which helped improve the clarity and quality of this paper. The authors did not receive any specific funding or support for this work and have no additional acknowledgements to report.

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

Received 5 December 2024; Revised 3 April 2025; Accepted 10 April 2025; Published online 25 September 2025

References

1. Geman R, Bensberg IR, Dozier W, Sholder SJ. 2023. Amended complaint from the plaintiffs in *Authors Guild v. OpenAI. Free Law Project*. www.courtlistener.com/docket/67810584/39/authors-guild-v-openai-inc
2. CourtListener. 2024. Authors Guild v. OpenAI Inc. (1: 23-cv-08292). www.courtlistener.com/docket/67810584/authors-guild-v-openai-inc/page=1
3. Patterson LR, Lindberg SW. 1991. *The nature of copyright: A law of users' rights*. Athens: University of Georgia Press. pp. 93–102
4. *Copyright Act of 1976*, 17 U. S. C. §§ 101-810
5. Seltzer LE. 1978. *Exemptions and Fair Use in Copyright: The Exclusive Rights Tensions in the 1976 Copyright Act*. Cambridge, MA, USA: Harvard University Press. pp. 19–20 doi: 10.4159/harvard.9780674433120
6. Gratz JC, Gass AM, Damle SV, Stillman AL. 2024. Defendants' response to complaint in *Authors Guild v. OpenAI. Free Law Project*. www.courtlistener.com/docket/67810584/75/authors-guild-v-openai-inc/
7. Buccafusco C, Masur JS. 2019. Intellectual property law and the promotion of welfare. *Research Handbook on the Economics of Intellectual Property Law*, eds. Depoorter B, Menell P. Vol. 1. Northampton: Edward Elgar Publishing. pp. 98–118. doi: 10.4337/9781789903997.00011
8. Kretschmer M, Kawohl F. 2004. The history and philosophy of copyright. In *Music and Copyright*. 2nd Edition. Edinburgh: Edinburgh University Press. pp. 21–53 doi: 10.1515/9781474468282-003
9. Rajan MTS. 2011. *Moral rights: Principles practice and new technology*. New York: Oxford University Press. pp. 7. doi: 10.1093/acprof:osobl/9780195390315.001.0001
10. Martin SM. 2002. The mythology of the public domain: Exploring the myths behind attacks on the duration of copyright protection. *Loyola of L. A. Law Review* 36(1):253–22
11. Boyle J. 2008. *The public domain: enclosing the commons of the mind*. New Haven, CT: Yale University Press. pp. 48–206 doi: 10.12987/9780300142754
12. Anonymous. 2024. Copyright Infringement/Vicarious Liability. *Business Torts Reporter* 36(6):128-31
13. Gass AM, Damle SV, Stillman AL, Gratz JC. 2024. OpenAI corporate disclosure statement from *Authors Guild v. OpenAI*. pp. 2–3. www.courtlistener.com/docket/67810584/73/authors-guild-v-openai-inc
14. OpenAI. 2025. *Our structure*. https://openai.com/our-structure/
15. Boyd E. 2023. *ChatGPT is now available in Azure OpenAI Service*. https://azure.microsoft.com/en-us/blog/chatgpt-is-now-available-in-azure-openai-service/#:~:text=Since%20ChatGPT%20was%20introduced%20late,helping%20with%20software%20programming%20questions
16. Moneycontrol. 2024. *Microsoft CEO Satya Nadella Exclusive Interview | AI-Led Business Will Add To $5T Indian Economy*. www.youtube.com/watchv=VP2O0AC9hn4
17. Kay G, Jackson S. 2024. The history of OpenAI, from the early days with Elon Musk to the ChatGPT maker being put on blast by Scarlett Johansson. www.businessinsider.com/history-of-openai-company-chatgpt-elon-musk-founded-2022-12#at-the-time-musk-said-that-ai-was-the-biggest-existential-threat-to-humanity-3
18. Brockman G. 2019. *Microsoft invests in and partners with OpenAI to support us building beneficial AGI*. https://openai.com/index/microsoft-invests-in-and-partners-with-openai/
19. OpenAI. 2023. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. https://openai.com/index/gpt-4/
20. Stollnitz B. 2023. *How GPT models work: for data scientists and ML engineers*. https://bea.stollnitz.com/blog/how-gpt-works-technical/
21. AI Basics. 2024. *In The Reference Shelf: New Developments in Artificial Intelligence*, ed. Wilson HW. Amenia, NY: Grey House Publishing. pp. 1–4
22. Atkinson-Abutridy J. 2024. *Large Language Models: Concepts, Techniques and Applications*. Boca Raton: CRC Press. pp. 1–74 doi: 10.1201/9781003517245
23. Kamath U, Graham K, Emara W. 2022. *Transformers for machine learning: a deep dive*. New York: Chapman and Hall/CRC. pp. 2–158 doi: 10.1201/9781003170082
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*. pp. 6000–10 doi: 10.48550/arXiv.1706.03762
25. Thakur K, Barker HG, Pathan AK. 2024. Artificial Intelligence and Large Language Models: An Introduction to the Technological Future. Boca Raton: CRC Press. pp. 91. doi: 10.1201/9781003474173
26. Wolfram S. 2023. *What Is ChatGPT Doing … and Why Does It Work?* https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/#:~:text=It%27s%20Just%20Adding%20One%20Word,text%20is%20remarkable%2C%20and%20unexpected
27. Foster D. 2019. *Generative deep learning: Teaching machines to paint, write, compose, and play*. Sebastopol, CA: O'Reilly Media, Inc. pp. 2–34
28. Rumelhart RE, Durbin R, Golden R, Chauvin Y. 1995. Backpropagation: The basic theory. In *Backpropagation: Theory, Architectures, and Applications*, eds. Chauvin Y, Rumelhart RE. New York: Psychology Press. pp. 7
29. Metz C. 2023. *The Secret Ingredient of ChatGPT Is Human Advice*. New York Times, September 25, 2023. www.nytimes.com/2023/09/25/technology/chatgpt-rlhf-human-tutors.html
30. Brooke R. 2023. *Fair Use Week 2023: Looking Back at Google Books Eight Years Later*. www.authorsalliance.org/2023/02/24/fair-use-week-2023-looking-back-at-google-books-eight-years-later
31. OpenAI. 2024. *How ChatGPT and our foundational models are developed*. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed
32. Martin GRR. 2002. *A game of thrones*. New York: Bantam Books. pp. 1. as quoted by ChatGPT
33. Frye N. 1957. *Anatomy of criticism: Four essays*. Princeton: Princeton University Pres. pp. 97
34. Kleon A. 2012. *Steal like an artist*. New York: Workman Publishing Company. pp. 7
35. Adobe. 2024. *Our approach to generative AI with Adobe Firefly*. www.adobe.com/ai/overview/firefly/gen-ai-approach.html

36. Authors Alliance. 2024. *What happens when your publisher licenses your work for AI training*. www.authorsalliance.org/2024/07/30/what-happens-when-your-publisher-licenses-your-work-for-ai-training/
37. Cambridge University Press. 2024. *Uncharted territory: AI and the Cambridge approach for academic book publishing*. www.cambridge.org/universitypress/about-us/news-and-blogs/uncharted-territory-ai-and-the-cambridge-approach-for-academic-book-publishing
38. OpenAI. 2024. *Terms of Use*. https://openai.com/policies/row-terms-of-use
39. MacKie-Mason J, Li H. 2023. *Re: Notice of inquiry ("NOI") and request for comments, Artificial Intelligence and Copyright, Docket No. 2023-6*. https://drive.google.com/file/d/1nNvynoekNREUhLA_h5AqE5HzVuTxymQm/view
40. Brinkhaus HO, Rajan K, Schaub J, Zielesny A, Steinbeck C. 2023. Open data and algorithms for open science in AI-driven molecular informatics. *Current Opinion in Structural Biology* 79:102542
41. xxx. 2024. *Artificial Intelligence Policy Act. Utah Code. Tit. 13, Ch. 72*. https://le.utah.gov/xcode/Title13/Chapter72/C13-72_2024050120240501.pdf
42. Goldstein P. 2003. *Copyright's highway: From Gutenberg to the celestial jukebox*. Revised Edition. Stanford: Stanford University Press. pp. 21