

BARTZ V. ANTHROPIC: FAIR USE IN BALANCING AI TECHNOLOGICAL DEVELOPMENT AND COPYRIGHT IN AI TRAINING DATA

Ifeoluwa A. Olubiyi*

Abstract

Access to copyright works is central to the training of Artificial Intelligence (AI) models, particularly, Large Language Models (LLMs). This paper reviews the application of the doctrine of fair use to AI training data by the US Court in Bartz et al. v. Anthropic in comparison to the earlier case of Thomson Reuters Enter. Centre GmbH v. Ross Intell. Inc and latter decision in Kadrey v. Meta Platforms, Inc. The paper argues that Bartz v. Anthropic may provide some guidance in balancing the interests of AI development and copyright owners by separating the issues of infringing AI output from AI training data. It also allows fair use for copyright protected AI training data only where the works were legally obtained which still gives room for compensation to copyright owners which ensures that copyright does not stifle technological advancement.

Keywords: Artificial Intelligence (AI), AI Training Data, Copyright, Fair Use, Fair Dealing, Text and Data Mining (TDM)

1. Introduction

In the last few years, there have been exponential developments in artificial intelligence (AI) technology. Generative AI (GenAI) can generate, modify or review documents, images, audio/music, videos, codes among others. The goal of AI is to enable machines to mimic human thoughts and behaviours including learning, reasoning and predicting.¹ If AI is to think and act as humans, then it must learn from humans how they think, see, understand, identify, interpret, process data among others. This means that AI needs to be trained on existing data produced by human beings or documented by humans.

Copyright is a significant intellectual property right that protects the artistic, literary and musical creation of humans as a product of their intellect, creativity and financial investments.² Copyright grants limited monopoly to copyright owners so they can control the access, use and exploitation of their copyright works.³ Authors of copyright works have certain moral rights including the right to be recognised in this capacity and to prevent derogatory or dishonourable modifications of their works.⁴ Copyright owners also generally have a right to be compensated financially for the uses of their works that do not fall within the limitations and exceptions recognised under the law.⁵

*LLB (OAU), BL, LLM (Maastricht), Ph.D (Ilorin),McArb; ORCID id: orcid.org/0000-0001-5328-5175. Alexander von Humboldt Research Fellow, Chair of Intellectual Property Law (Prof. Dr. Christoph Ann), School of Management, Technical University of Munich, TUM, Germany. ifeoluwa.olubiyi@tum.de; Professor, Department of Private and Business Law, College of Law, Afe Babalola University Ado-Ekiti (ABUAD), Nigeria. olubiyia@abuad.edu.ng

¹Yongjun Xu & Ors, 'Artificial Intelligence: A Powerful Paradigm for Scientific Research' Innovation (Camb) (2021) 2 (4):100179. Doi: 10.1016/j.xinn.2021.100179.

² John Hutson, 'The Evolving Role of Copyright Law in the Age of AI-Generated Works' (2024) 2(4) Journal of Digital Technologies and Law 886-914 DOI: <https://doi.org/10.21202/jdtl.2024.43>; Hafiz Gaffar & Saleh Albarashdi, 'Copyright Protection for AI-Generated Works: Exploring Originality and Ownership in a Digital Landscape' 2025;15(1) Asian Journal of International Law 23-46. DOI:10.1017/S2044251323000735.

³ Christian Peukert & Margaritha Windisch, 'The Economics of Copyright in the Digital Age' (2025) 39 Journal of Economic Surveys 877-903. <https://doi.org/10.1111/joes.12632>.

⁴ Nigerian Copyright Act 2022, s.14.

⁵ Ibid, s.9-13.

AI needs to be trained on large data sets to be effective, many of which are protected by intellectual property rights including copyright.⁶ The situation of copyright owners is unique because the issue is not merely that AI has been trained on their works without authorisation or compensation. The more significant consequence is that GenAI is able to reproduce their works wholly or a significant proportion and distribute same to an unlimited number of people.⁷ In addition, AI can also modify their works even in ways that are considered derogatory or contrary to their reputation.⁸ Then comes also the palpable concerns that AI not only creates a competition for copyright owners but also has the potential to replace them in the creative and innovative sector.⁹ On the part of AI companies, arguments have included that allowing the use of copyright works is essential for the purpose of promoting technological development.¹⁰ Other arguments advanced include that the use of these works as AI training data amounts to a limitation or exception to copyright under fair use or fair dealing or other limitations to copyright law such as text and data mining as available in some jurisdictions such as the EU Copyright Directive.¹¹

In the absence of clear legislative guidance, the court has remained a key forum for navigating the legal regime for the intersection between AI and intellectual property rights. The case of *Bartz et al. v. Anthropic PBC*¹² is significant because the court, in applying the doctrine of fair use to AI training data, distinguished between AI training and AI output as well as between legally acquired copyright works as training data and illegally acquired ones. This provided an avenue for the court to balance the interests of technological development while not shutting the door against compensation for right owners. So far, it is also the largest copyright settlement case in the United States of America (USA). Therefore, it plays a pivotal role in providing some guidance on the interplay of copyright and the use of copyrighted materials as AI training data. This paper shall briefly state the facts of this case and proceeds to critically analyse the reasoning of the court. Comparison shall be made to the earlier case of *Thomas Reuters v. Ross Intelligence*¹³ and the subsequent case of *Kadrey v. Meta*¹⁴ in analysing how the court has approached the application of fair use to AI training data.

2. *Bartz v. Anthropic*

Sometime in 2024, the claimants, Andrea Bartz, Charles Graeber, and Kirk Wallace Johnson, who are authors sued Anthropic, the AI company behind the Claude AI language models for using their

⁶ David W. Opderbeck 'Copyright in AI Training Data: A Human-Centered Approach' (2024) 76 Oklahoma Law Review 951-1023.

⁷ Antonia Karamolegkou & Ors, 'Copyright Violations and Large Language Models' (2023) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing 7403–7412; Felix B Mueller & Ors, 'LLMs and Memorization: On Quality and Specificity of Copyright Compliance' (2024) 7(1) Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 984-996 DOI: <https://doi.org/10.1609/aies.v7i1.31697>.

⁸ Felipe R. Moreno, 'Generative AI and Deepfakes: a Human Rights Approach to Tackling Harmful Content' (2024) 38(3) International Review of Law, Computers & Technology 297–326. DOI: <https://doi.org/10.1080/13600869.2024.2324540>

⁹ Bender Stuart, 'Generative-AI, the Media Industries, and the Disappearance of Human Creative Labour' (2025) 26 (2) Media Practice and Education 200–217. Doi:10.1080/25741136.2024.2355597; Sabine Jacques & Mathew Flynn, 'Protecting Human Creativity in AI-Generated Music with the Introduction of an AI-Royalty Fund' (2024) 73 (12) GRUR International 1137–1149, <https://doi.org/10.1093/grurint/ikae134>

¹⁰ Eleonora Rosati, 'Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity' (2019) 27 Asia Pacific Law Review 198; Mark A Lemley & Bryan Casey, 'Fair Learning' (2021) 99 Texas Law Review 743.

¹¹ Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC Directive (Directive 2019/790), articles 3&4

¹² 3:24-cv-05417, (N.D. Cal. filed Aug. 19, 2024)

¹³ 765 F. Supp. 3d 382, 398 (D. Del. 2025).

¹⁴ 3:23-cv-03417 (N.D. Cal. filed July 7, 2023).

copyrighted works to train Claude without authorisation or compensation. The defendants had attempted to amass a central library of “all the books in the world” to retain “forever.”¹⁵ The claimants argued that the actions of the defendant constituted direct copyright infringement and unlawful circumvention.¹⁶ Anthropic had downloaded over 7 million copyrighted books from ‘shadow libraries’ like Library Genesis (LibGen) and Pirate Library Mirror (PiLiMi) as AI training data. However, the company also legally bought and scanned millions more books and copyrighted works for use as AI training data including some of those it had illegally obtained earlier. The defendant argued that its use of publicly available data including copyright works was a transformative use of the works which amounted to fair use under section 107 of the US Copyright Act.

In arriving at its decision, the court separated issues around the claim to reflect the use of copyright works as AI training data generally, and the question of the mode of acquisition or access to the copyright works. On the first issue, the court generally ruled that the use of legally acquired copyright works to train AI was a ‘quintessentially transformative’ use of the work and falls within the scope of fair use as a copyright exception under the US Copyright Act.

However, illegally accessing copyright works such as downloading copies from shadow libraries or pirated copies and keeping them did not amount to fair use and the court suggested this would amount to infringement. In August 2025, Judge Alsup certified the case as a class action, defining the class to include ‘all beneficial or legal copyright owners of the exclusive right to reproduce copies of any book’ in the pirated datasets, with the condition that those works must have an ISBN/ASIN and also be registered with the U.S. Copyright Office.¹⁷ Hence, the claimants became representatives of not only authors but also publishers, estates, and anyone else with reproduction rights in the protected works. With the availability of statutory damages in the American legal system, which could be up to \$150,000 per work, Anthropic was potentially liable to damages in billions of dollars to copyright owners. The parties thereafter agreed on a settlement of \$1.5 billion with \$3,000 dollars per copyright work identified to have been infringed. It must be noted that the scope of this settlement is limited to only the past actions of Anthropic and does not set a precedence or licence for subsequent AI training data or output that may be infringing on copyright.

3. Fair Use: Analysis of Factors and Comments

Fair use is a general exception to copyright which serves as a limitation and exception to the right of the copyright owner to balance the interests of the copyright owner to control the use of copyright works and the interests of the public to have access to information.¹⁸ While respecting the rights of copyright owners, the public still needs access to information, data, materials and knowledge in copyright protected works to serve public interests including inspiration, raw material for creativity, educational, cultural, scientific and technological advancements.¹⁹ The concept of

¹⁵ *Bartz v. Anthropic PBC*, 3:24-cv-05417, (N.D. Cal.), Doc 231 Order on Fair Use (Filed 06/23/25) [subsequently referred to as ‘Bartz Order on Fair Use’] 1.

¹⁶ Copyright Act (17 USC), ss 106, 107, 504; Digital Millennium Copyright Act (DMCA), s 1202.

¹⁷ *Andrea Bartz v. Anthropic PBC* (Memorandum Opinion on Preliminary Approval of Class Action Settlement) Case 3:24-cv-05417-WHA Document 437 Filed 10/17/25.

¹⁸ Desmond O Oriakhogba & Ifeoluwa A Olubiya, *Intellectual Property Law in Nigeria: Emerging Trends, Theories and Practice* (2nd ed, Paclerd Press) 256.

¹⁹ Christophe Geiger & Bernd J. Jütte, ‘Copyright as an Access Right: Concretizing Positive Obligations for Rightholders to Ensure the Exercise of User Rights’ (2024) 73 (11) *GRUR International*, 1019–1035, <https://doi.org/10.1093/grurint/ikae130>; WIPO Standing Committee on Copyright and Related Rights, ‘Scoping Study on Copyright Limitations and Exceptions for Educational and Research Activities in the Digital Age’ (WIPO SCCR/44/4, 2023) https://www.wipo.int/edocs/mdocs/copyright/en/sccr_44/sccr_44_4.pdf accessed 11 March 2026.

fair use or fair dealing is common to many common law jurisdictions including the USA and Nigeria.²⁰

The concept is quite flexible, and the law allows the judge to decide on a case-by-case basis allowing the court to exercise its discretion in administering fair use.²¹ Its strength is in its flexibility and adaptability as the term is not strictly defined in the law. The law permits the fair use of a work including the reproduction of a copyright work for the purpose of criticism, comment, news reporting, teaching, scholarship, or research without the authorisation of the copyright owner.²² This means that what amounts to fair use is not limited to only these expressly stipulated uses. The law only provides four factors to guide the court in arriving at its decision. These 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.'²³

On the first factor of the purpose and character of use, the court considered Anthropic's use of the copyrighted works as 'exceedingly transformative'²⁴ as the AI technology is among 'the most transformative many of us will see in our lifetimes'.²⁵ The works were not reproduced or distributed in any recognisable way, but rather abstracted into statistical representations that enabled language prediction capabilities. Hence, converting legally purchased physical books into digital copies for training was deemed acceptable, provided they were not creating new works or redistributing existing copies. 'In short, the purpose and character of using copyrighted works to train Large Language Models (LLMs) to generate new text was quintessentially transformative. Like any reader aspiring to be a writer, Anthropic's LLMs trained upon works not to race ahead and replicate or supplant them — but to turn a hard corner and create something different.'²⁶

Copyright Alliance²⁷ has criticised this interpretation of 'transformative use' for not applying the US Supreme Court's central consideration for a legal analysis of 'transformative use' which is whether the use "merely supersedes the objects of the original creation . . . (supplanting the original), or instead adds something new, with a further purpose or different character" as decided in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*.²⁸ Indeed the court in *Bartz* did not allude to this legal analysis and rather just concluded that AI is 'the most transformative many of us will see in our lifetime.' Nevertheless, even if this legal analysis was done, it can be argued that the use of copyright works in training AI actually 'adds something new' and is with 'a further purpose or different character,' as AI training data which is different from the original copyright works. Hence, using this legal analysis, it is possible that the court may arrive at the same conclusion.

²⁰ US Copyright Act, s.107; Nigerian Copyright Act 2022, s.20 (1). It must be noted that the concept of fair dealing in Nigeria is similar to fair use as the Nigerian Act also now contains these four guiding factors under the US fair use doctrine.

²¹ Glynn S. Lunney, Jr 'Transforming Fair Use' (2024)14 (1) NYU Journal of Intellectual Property and Entertainment Law 169-242.

²² US Copyright Act, s.107.

²³ Ibid.

²⁴ Bartz Order on Fair Use, 9.

²⁵ Ibid, 30.

²⁶ Ibid, 13-14.

²⁷ Kevin Madigan 'Transformative Use Analysis in Bartz v. Anthropic AI Case Marred by Fatal Flaws' <https://copyrightalliance.org/bartz-anthropic-ai-case-flaws/> accessed 9 March 2026.

²⁸ *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508 (2023).

More so, the U.S. Copyright Office in its report on Copyright and AI²⁹ and on the analysis of fair use noted that 'The use of a model may share the purpose and character of the underlying copyrighted works without producing substantially similar content. Where a model is trained on specific types of works in order to produce content that shares the purpose of appealing to a particular audience, that use is, at best, modestly transformative.' The report therefore acknowledges that using copyright works for training may be transformative, even if modestly, especially where the output is not substantially similar. Here also, we see a separation between AI training data and AI output in determining fair use.

The second factor for determining fair use is the nature of the copyrighted work. For instance, more protection is likely to be given to unpublished works than published works. Expressive, creative, fictional and fantasy works are also more protected than works of facts. In this instance, it was clear that the copyright works contained expressive elements and were chosen for this purpose. This factor therefore weighed against fair use.³⁰

With regards to the amount and substantiality of the work copied in relation to the original copyright work, the court agreed that the whole work has been ingested, reproduced, copied or used as AI training data. However, what matters is not just what was copied but what was made accessible to the public.³¹ Hence, the court clarified that this case was about the use of the works as AI training data; it was not about the AI output and so in this instance, no output from the model replicated those works verbatim. The court mentioned that if the claim was about AI output that reproduced the copyright works, there may have been a different analysis of question of infringement.³²

In assessing the last factor on the effect of use on the market for the original copyright work, the claimant argued that training LLMs displace or will displace an emerging market for licensing their works as AI training data, while the defendant argued that the transactional cost of this would exceed the benefits and AI companies may cease dealing with rightsholders or developing the technology. The court held that, in this instance, the plaintiff failed to show any market harm, either to the original work or any derivative AI training licensing market, which was still largely speculative at the time of the court action. The court deemed training AI models on books as a new, transformative purpose that does not necessarily replace the market for the original works, thus protected under fair use.

Of note is the statement of the court that such a licensing market 'is not one the Copyright Act entitles the authors to exploit'.³³ This is not necessarily so. Copyright gives the copyright owners wide rights which may be interpreted even across technological developments. A market of copyright licensing for AI training data does not expressly need to be contained in the Act before copyright owners can take advantage of it. Taking a cue from previous court decisions, the validity and protection of intellectual property rights have been maintained or expanded even within the realms of emerging technologies such as the Metaverse or as Non-Fungible Tokens.³⁴ The rights were naturally extended to cover new medium of exploitation of IPR. This should not be different for AI technology. The law cannot always feasibly be amended fast enough to take emerging technologies into consideration; often the courts continue to provide guidance until the law catches up legislatively.

²⁹ US Copyright Office, 'Copyright and Artificial Intelligence (Part 3: Generative AI Training) (May 2025) 46 <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>.

³⁰ Bartz Order on Fair Use, 24

³¹ Ibid, 25. The referred to the decision in *Authors Guild v. Google, Inc.*, 804 F.3d 202, 222 (2d Cir. 2015)

³² Bartz Order on Fair Use, 7, 11-12

³³ Ibid, 28

³⁴ *Hermès v Rothschild* 22-cv-384 (JSR); *Nike, Inc. v. StockX LLC* 1:22-cv-00983-VEC;

Juventus Football Club S.p.A., v Blockeras s.r.l (Docket No. 32072/2022, Court of Rome IP Chamber, 20/07/2022)

The industry is also providing solutions by negotiating licences with respect to AI use or training data such as the one for images or personality rights. For instance, the Screen Actors Guild - American Federation of Television and Radio Artists (SAG-AFTRA) has entered into licensing agreements for the use of its members' voices with AI companies. The SAG-AFTRA & Replica Studios (Jan 2024) Agreement allowed its union members to license digital replicas of their voices for use in video games and interactive media and the SAG-AFTRA & Narrativ (Aug 2024) Agreement lets performers license their voices for use in audio advertisements.³⁵ In 2023, New York Times sued OpenAI and Microsoft, for copyright infringement by using millions of its articles as AI Training data.³⁶ By May 2025, The New York Times announced that it had entered an AI licensing deal with Amazon.³⁷ Other news outlets, such as Axel Springer, Condé Nast, The Washington Post and News Corp have also entered into licensing agreements for the use of their materials.³⁸

In *Bartz*'s case, the claimants had argued that training LLMs will result in an explosion of works competing with their works. The court, though agreeing with this possibility, was not persuaded by this argument and felt that this was no different from training school pupils to write well using copyright works and arguing that this will lead to an explosion of competing works. It stated that 'this is not the kind of competitive or creative displacement that concerns the Copyright Act. The Act seeks to advance original works of authorship, not to protect authors against competition.'³⁹ While it is true that AI LLMs are being equipped to compete with copyright owners, one must bear in mind that the issue before the court in this instance, was a question of the use of copyright works as AI training data and not that AI outputs are infringing on copyright. This distinction was clearly made by the court. Hence, this decision does not preclude claims of copyright infringement by AI outputs or generated works some of which have been litigated and the court held that they are infringing.⁴⁰ However, the fears of copyright owners are valid considering that unlike the analogy of school pupils used, AI can produce more 'perfect' works in few seconds and disseminate them to a potential worldwide audience. Nevertheless, the focus of this case is AI training data and not output. Hence, this may be the reason the court did not agree with it.

Bartz's case makes it clear that AI training data should be based on legally acquired copyright works while ensuring that measures are put in place to prevent unauthorised access or distribution of the copies of the works. Where necessary, access should be for only the stipulated purpose and

³⁵ SAG-AFTRA 'SAG-AFTRA and Replica Studios Introduce Groundbreaking AI Voice Agreement at CES' <https://www.sagaftra.org/sag-aftra-and-replica-studios-introduce-groundbreaking-ai-voice-agreement-ces> accessed 10 March 2026; SAG-AFTRA, 'SAG-AFTRA and Narrativ Announce New Agreement' <https://www.sagaftra.org/sag-aftra-and-narrativ-announce-new-agreement> accessed 10 March 2026. Replica Studios is an AI voice technology company and Narrativ is an online marketplace where brands can create audio ads using artificial intelligence.

³⁶ *The New York Times Company v. Microsoft Corporation*, 1:23-cv-11195, (S.D.N.Y.).

³⁷ Michael M. Grynbaum & Cade Metz, 'The Times and Amazon Announce an A.I. Licensing Deal' (29 May, 2025) <https://www.nytimes.com/2025/05/29/business/media/new-york-times-amazon-ai-licensing.html> accessed 10 March 2026.

³⁸ Benjamin Mullin, 'Inside the News Industry's Uneasy Negotiations with OpenAI' (29 December 2023) <https://www.nytimes.com/2023/12/29/business/media/media-openai-chatgpt.html> accessed 10 March 2026.

³⁹ *Bartz* Order on Fair Use, 28

⁴⁰ In *GEMA v. OpenAI* (Landgericht München I (Munich Regional Court I), Judgment of 11 November 2025 — 42 O 14139/24): The Munich Regional Court ruled that ChatGPT's output of song lyrics constituted copyright infringement; *Getty Images v. Stability AI* ([2025] EWHC 2863 (Ch). Case No: IL-2023-000007): A UK High Court found trademark infringement because earlier versions of Stable Diffusion generated images containing distorted versions of Getty's 'iStock' and 'Getty Images' watermarks; *Andersen v. Stability AI* (Case No: 3:23-cv-00201-WHO (N.D. Cal.) Ongoing): A U.S. court allowed direct infringement claims to proceed based on evidence that AI systems could 'recreate' specific copyrighted training images through precise prompts.

not retained longer than it is necessary to achieve the purpose of the AI training. AI companies may therefore consider a licensing scheme to be on a safe side.

The court in *Bartz* distinguished its decision from the earlier case of *Thomson Reuters Enter. Centre GmbH v. Ross Intell. Inc.*⁴¹ In this case, Thomson Reuters, owner of Westlaw legal research platform, sued Ross Intelligence, an AI-driven legal research startup. Westlaw features 'headnotes,' which are editorially compiled summaries of legal points from court judgments, along with a proprietary 'Key Number System' for organising them. The defendant, intending to build a competitor to Westlaw, asked to license the claimant's content which request the claimant refused because of direct competition. Subsequently, the defendant hired a third party, LegalEase, to create 'Bulk Memos', a collection of legal questions and answers designed to train their AI to identify relevant case law. The claimant sued claiming that LegalEase created these memos by using Westlaw's copyrighted headnotes. The court ruled that Westlaw's headnotes possess sufficient 'minimal creativity' as it involved distilling and synthesising legal points which are copyrightable although the underlying judicial opinions are part of the public domain.

The court found that Ross copied 2,243 headnotes in creating their training data and rejected Ross's fair use defence. In deciding on fair use, the court held that on the first factor of 'purpose and use' the defendant's use was not transformative because they used the headnotes for the same purpose as Westlaw (facilitating legal research) to build a competing product. Also, on the fourth condition of effect on the market, the court found that the defendant's use negatively impacted the potential market for licensing Westlaw's content for AI training. The court in *Bartz* emphasised twice while discussing 'purpose and character' that what was trained was 'not generative AI (AI that writes new content itself). Rather, what was trained — using a proprietary system for finding court opinions in response to a given legal topic — was a competing AI tool for finding court opinions in response to a given legal topic. That was not transformative'.⁴²

It is also important to look at the case of *Kadrey v. Meta Platforms, Inc.*⁴³ decided just 2 days after the decision in *Bartz*. Thirteen authors, including Richard Kadrey, sued Meta, alleging that Meta infringed their copyright by training its large language models on their books. Some of the books were allegedly acquired from so-called shadow or pirate libraries like Library Genesis or Bibliotik. *Kadrey's* decision may have confirmed that the approach of the US court is to permit the use of copyrighted works as AI training data under the fair use exception. In this case, the court ruled that Meta's use of copyrighted books to train its AI models constitutes fair use, primarily because the plaintiffs failed to provide evidence of market harm. The court found the training process 'highly transformative' and that copying entire works was necessary for creating functional LLMs. It found that such a use repurposed the books into a tool that could generate summaries, stories, or code which is a fundamentally different use from simply reading the books. Furthermore, it held that copying works wholly was necessary to achieve the transformative purpose of training large language models; hence, the amount copied was reasonable in this context. Finally, the court held that the plaintiffs offered no evidence that LLMs' outputs acted as substitutes for their books or otherwise harmed sales, although it acknowledged the theoretical risk of market dilution, 'where AI-generated content could saturate the market with derivative or style-based works'.

Unfortunately, in determining its finding of fair use, the court in *Kadrey* did not consider the mode of access or acquisition of the copyright materials used as training data as was done in *Bartz*. It did not focus on whether the texts came from legitimate publishers or shadow libraries. It is, however, worthy of note to consider the aspect of retention of the copyright materials for a substantial period which may be another distinguishing factor of the two cases. Meta used

⁴¹ 765 F. Supp. 3d 382, 398 (D. Del. 2025) (Judge Stephanos Bibas), appeal docketed, No. 25-8018 (3d Cir. Apr. 14, 2025).

⁴² *Bartz* Order on Fair Use, 13.

⁴³ No. 3:23-cv-03417 (N.D. Cal. filed July 7, 2023).

'torrenting' to assemble copies of the books and post-trained its models to prevent them from memorising and outputting certain text from their training data, including copyrighted material. The Copyright Office in its Copyright and AI report⁴⁴ seems to also agree that the retention and memorisation of copyright works by LLMs may be put into consideration in considering fair use. The report states that 'Whether a model's weights implicate the reproduction or derivative work rights turns on whether the model has retained or memorized substantial protectable expression from the work(s) at issue... the use of those works in preparing a training dataset and training a model implicates the reproduction right, but copying the resulting weights will only infringe where there is substantial similarity.' This is another point of distinction in *Bartz's* case where Anthropic did not only train its LLM using the copyright works but also sought to maintain a library 'forever' which potentially means it could be used beyond training, and the works were being retained beyond what is necessary. While the training itself was fair use, the court ruled that maintaining this separate repository was not protected.

It is settled that the flexibility of fair use makes it suitable to be applied to instances that may not have been mentioned in the law but falls within the scope of the expressly mentioned uses and generally what is fair in the circumstances. Hence, it is possible to apply it to modern uses of copyright works as necessitated by technological developments such as computer technology, the internet and AI in this instance.⁴⁵ Katz opines that this doctrine will permit unauthorised use of copyright works in a manner that promotes public interest in the encouragement of dissemination of works of arts when the use does not seriously jeopardise the copyright owner's opportunity to obtain a just reward.⁴⁶

Now we look at the concluding remarks by the US Copyright Office in its analysis of Fair Use and AI training Data in its report:

Various uses of copyrighted works in AI training are likely to be transformative. The extent to which they are fair, however, will depend on what works were used, from what source, for what purpose, and with what controls on the outputs—all of which can affect the market. When a model is deployed for purposes such as analysis or research—the types of uses that are critical to international competitiveness—the outputs are unlikely to substitute for expressive works used in training. But making commercial use of vast troves of copyrighted works to produce expressive content that competes with them in existing markets, especially where this is accomplished through illegal access, goes beyond established fair use boundaries.⁴⁷

The above indicates that the use of copyright works as AI training data may be transformative and permitted by fair use. A rejection of fair use may be tied to production of similar works to compete in existing markets and illegal access to those copyright works. Therefore, infringing AI outputs and illegal access to AI training data may be the instances when fair use may not be available for AI companies.⁴⁸ Furthermore, it is only fair based on the principle of exhaustion or first sale doctrine of copyright that where a copyright work has been legally obtained and the copyright owner compensated for its use, in this case, as AI training data, then the user should be allowed to

⁴⁴ US Copyright Office, 'Copyright and Artificial Intelligence (Part 3: Generative AI Training) (May 2025) p.30 <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>.

⁴⁵ Desmond O Oriakhogba & Ifeolwa A Olubiya, *Intellectual Property Law in Nigeria: Emerging Trends, Theories and Practice* (2nd ed, Paclerd Press) 258.

⁴⁶ Ariel Katz 'Fair Use 2.0: The Rebirth of Fair dealing in Canada' in M Geist (ed.) *The Copyright Pentology: How the Supreme Court of Canada shook the Foundations of Canadian Copyright Law* (2013) 93-156

⁴⁷ US Copyright Office, 'Copyright and Artificial Intelligence (Part 3: Generative AI Training) (May 2025) 107 <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>.

⁴⁸ Peter Henderson & Ors, 'Foundation Models and Fair Use' (2023) 24 Journal of Machine Learning Research 1-79

make use of the copyright work for this purpose and the copyright owner may not seek additional compensation.⁴⁹ Although in *Bartz*'s case, the court did not consider the question of exhaustion of right. Rather, based on the requirement of 'transformative use' it held Anthropic's format-change of its legally acquired print copies to digital copies was transformative under fair use factor one.⁵⁰ Hence, a mere format change for easing storage and searchability without redistributing was fair. It is important to state that some jurisdictions that do not have fair use have the copyright exception of text and data mining (TDM) which allows copyright works to be used and analysed and in this case as AI training data without the author's consent. TDM is any automated analytical technique that analyses text and data in digital form to generate information, including patterns, trends, and correlations.⁵¹ Training of AI models involves the automated analysis of data to create new information (LLM), therefore, TDM exception is useful for AI training.

For example, the EU Copyright in the Digital Single Market⁵² contains the TDM exception for scientific research under article 3 and a general TDM exception in article 4. Article 3 provides for a mandatory TDM exception for scientific research by research organisations and cultural heritage institutions. It is limited to non-commercial research and sometimes public-private partnership research. The entities must however have lawful access to the data or content being mined. Article 4 provides for a general TDM exception which can be for any purpose, including commercial uses, for anyone with lawful access to works including contents that are freely available online. However, rightsholders can opt-out of this exception to prevent TDM, provided they do so in a machine-readable manner. Generally, copies of works made for TDM must be stored securely and can only be retained for as long as necessary for the purpose of the mining. The exceptions also do not permit sharing or publication of the copyrighted material itself, unless authorised.

This requirements of lawful access to enjoy TDM exception, storage and use of the data for the stipulated purposes and retention only for as long as necessary align with the decision of the court in *Bartz*. It appears that jurisdictions with a TDM already have an exception to prevent stifling the development of AI technology by allowing the use of copyright works as AI training data. Common law jurisdictions with no such provision may therefore tow the path of *Bartz* to allow the use of copyright works for AI training but require a legitimate acquisition or access to such works, use, storage and retention only as necessary or required for the purpose of the AI training.

As expected, not everyone agrees with the decision in *Bartz* and application of fair use exception to AI Training data. For Nam, fair use was originally designed as an exceptional system for 'little users,' but AI developers and Big Tech companies have abused this doctrine in the name of 'public interest' which leads to the 'paradox of fair use'. Nam considers TDM 'an expanded and advanced version of fair use' and these two concepts raise concerns that creative works by humans may be reduced to mere 'firewood' for advancing AI performance.⁵³

Reinfeld & Gaon argue that the last factor of market effect should not simply be glossed over by the court in analysing fair use neither does the possibility of rightholders opting out under the TDM exception solve the challenge.⁵⁴ For the authors, the relevant question is not whether AI generated

⁴⁹ Kanchana Kariyawasam, Royal Raj Subburaj, Importance of the doctrine of digital exhaustion in copyright law, *International Journal of Law and Information Technology*, Volume 33, 2025, eaaf009, <https://doi.org/10.1093/ijlit/eaaf009>

⁵⁰ *Bartz* Order on Fair Use, p.17-18

⁵¹ Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC Directive (Directive 2019/790) art 2(2)

⁵² Directive (Directive 2019/790)

⁵³ Hyung Doo Nam, The Paradox of Fair Use – A Giant on a Seesaw: Restoring the Balance in Copyright and Media-Specific Issues' (2026) 2 GRUR International, 1 ikag001, <https://doi.org/10.1093/grurint/ikag001>

⁵⁴ Yuval Reinfeld & Aviv Gaon, 'Why IP's AI Debate Is Really About Market Power'(2026) 57 IIC 1–4 DOI:<https://doi.org/10.1007/s40319-026-01678-w>.

outputs replaces the original work, but whether large-scale AI training using copyrighted works contributes to the creation of a 'general-purpose substitute engine', that can fill the relevant market with competing outputs at little marginal cost and consequently erode both demand for future works and the creator's bargaining power in the long run. Hence, effect on the market is not only an abstract concept but where the crux of the disputes lie. A close examination reveals that the debates are often about who controls the market for data, information, creativity and innovation. With an unregulated access, the power dynamics begins to shift from rightholders to AI companies. However, despite these criticisms, the case of *Bartz* creates a balance in that it does not obliterate or refuse to recognise the rights of copyright owners. Illegal acquisition of copyright works was held to be infringing and the decision only applied fair use to copyright materials which were legally obtained. This means that before AI companies can use such works, they must have compensated the copyright owners for its use. This decision does not therefore obliterate the licensing market for copyright works as AI training data. As can be seen in examples of recent industry practice above, AI companies are negotiating licences with groups or representatives of copyright owners. This system is not perfect as issues will surround compensation for copyright owners who are not a part of collective management system. However, addressing that issue is beyond the scope of this paper.

4. Conclusion

While IPR cannot be used to stifle technological developments, fair use, fair dealing, and TDM may be a tool to help balance the interest of technological advancement by allowing copyrighted works to be used as AI training data. However, this cannot absolve the AI companies for liability for infringing outputs. This distinction may be suitable for balancing these interests and affirm the *Betamax* doctrine that copyright or intellectual property rights cannot be used to stifle technological advancements especially where there are legitimate non-infringing uses of a technology.⁵⁵ One of the rationales of intellectual property rights is to promote scientific and technological advancement. The contours of copyright must continue to be adjusted to accommodate this goal while also protecting the interests of copyright owners.

For fair use to avail AI companies, copyright works should have been legally obtained which ensures compensation to copyright owners. This position still leaves room for AI training licensing agreements. More so, such works may not be redistributed and should be retained only as necessary for the intended purposes. More so, AI companies must put technological measures in place to prevent infringing AI outputs, otherwise, they will be held liable for such acts of infringement. Where AI output infringes on protected works by producing substantially similar or identical works, then such a use will no longer be transformative but competing in purpose, use and market with the original copyright work. In such instances, fair use should not absolve a defendant of liability.

⁵⁵ *Sony Corp. of America v. Universal City Studios, Inc.*, 464 U.S. 417 (1984) (*Betamax* case).