Intellectual Property and Artificial Intelligence
19th February 2025



Chidumebi Nwosu[1]

THE LEGALITY OF USING PUBLICLY AVAILABLE INFORMATION FOR TRAINING ARTIFICIAL INTELLIGENCE MODELS

1 INTRODUCTION

Artificial Intelligence (AI) has reshaped the way humans interact with technology, promoting advancements in natural language processing, image recognition, and decision-making systems. At the centre of this development are large language models (LLMs), such as OpenAI's GPT-4, Claude Ai, etc, which are trained on vast datasets drawn from multiple sources. A large portion of the data used by these models are obtained from publicly available information, including websites, blogs, and social media platforms. The latest models of GPT are trained on trillions of words, a dataset so big that it would be "the equivalent of a Microsoft Word document that is over 3.7 billion pages long".[2] While this practice may seem to promote innovation, it equally raises certain legal and ethical concerns.

On one hand, proponents may argue that training of AI with publicly available information does not infringe on any intellectual property rights as most of the data being relied upon is already available in the public space. On another hand, it could be argued that the context in which the data was publicized may be important in determining what is appropriate in terms of its use and application. Moreso, the

---

[1] **Chidumebi Nwosu, NYSC Associate**, Intellectual Property and Technology Department, S.P.A Ajibade& Co., Lagos, Nigeria.

[2] See, 'Audrey Pope NYT v. OpenAI: The Times's About-Face' available at <https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/> accessed on 7th February 2025.

accessibility of a protected work does not diminish the restriction on the use of such work.

The use of publicly available information for training Artificial Intelligence models thus raises significant questions. Can copyrighted materials be used in training AIs without explicit permission? Does the use of publicly available information on websites infringe on privacy laws or terms of service agreements? Does the use of these copyrighted material for AI training fall under the fair use exception? These questions highlight the struggle between the need for innovation and that of protecting the rights of authors of creative works alongside personal data of individuals. This article seeks to explore the frameworks that govern such practices, the ethical implications, the challenges and opportunities for stakeholders in navigating this novel area.

2 DIFFERENCES BETWEEN PUBLICLY AVAILABLE INFORMATION AND INFORMATION IN THE PUBLIC DOMAIN.

2.1 Publicly Available Information

Publicly available information is any data that can be accessed without any special permission or legal authority, encompassing, government records, news and media content, business and financial data, social media and online content, academic and research publications etc.[3] Despite its accessibility, this type of content is still typically protected by copyright law.[4] This means that although anyone can view or read the content, using it in other contexts (like training an AI model) must be done with the permission of the copyright holder or within the context of what may constitute an exception under the relevant legislations.[5]

2.2 Information in the Public Domain

In contrast, information in the public domain consists of works free from copyright restrictions. Ideas, pictures, sounds, discoveries, facts, and texts that are not covered

---

3 See, Greg Mcfee 'Understanding Publicly Available Information: Definition, Uses, and Legal Restrictions' available at <https://electronmagazine.com/understanding-publicly-available-information-definition-uses-and-legal-restrictions/> accessed on 7th February 2025.

4 See, Beth Johnson 'Can I Use Publicly Available Data for Research or Projects Without the Risk of Copyright Infringement?' available at <https://www.copyright.com/blog/can-i-use-publicly-available-data-for-research-or-projects-without-the-risk-of-copyright-infringement/> accessed on 8th February 2025.

5 See, Keith Kupferschmid 'Insights from Court Orders in AI Copyright Infringement Cases' available at <https://copyrightalliance.org/ai-copyright-infringement-cases-insights/> accessed on 9th February 2025.

by intellectual property rights and are available for anyone to use or expand upon are all considered to be in the public domain.[6]

Works enter the public domain either because their copyright has expired, the creator has intentionally relinquished rights, or they were never eligible for copyright protection. Classic literature, certain government documents, and some academic works fall into this category. Once in the public domain, anyone is free to use, modify, and distribute the material without seeking permission.

The key distinction between information in the public domain and publicly available information is therefore not in their accessibility, since both can be easily accessed, but in the restrictions on their use. Publicly available information may remain subject to copyright and data privacy laws, meaning its appropriation requires permission or must fall within specific exceptions. In contrast, information in the public domain is free from such restrictions, allowing anyone to use, modify, and distribute it without seeking permission.

3 LEGAL FRAMEWORK GOVERNING AI TRAINING

While there are no specific laws governing Artificial Intelligence training in most jurisdictions globally and particularly in Nigeria, the National AI Strategy (NAIS) provides a foundational framework that could influence future regulations on AI training, the draft was made public by the Federal Ministry of Communications, Innovation, and Digital Economy in August 2024 with drafting assistance from private organizations, the National Information Technology Development Agency (NITDA) and the National Centre for Artificial Intelligence and Robotics (NCAIR).[7]

A core focus of the NAIS is ensuring that AI development aligns with ethical principles such as fairness, transparency, accountability, privacy, and human well-being. This includes holistic data governance standards that adhere to the Nigeria Data Protection Act (NDPA), which could shape guidelines on how training data is collected, processed, and used. The strategy introduces five strategic pillars which includes;

i. Building Foundational AI Infrastructure,

ii. Building and Sustaining a World-class AI Ecosystem

---

6 See, Duke Law Center for the Study of the Public Domain 'Public Domain Frequently Asked Questions' available at <https://web.law.duke.edu/cspd/publicdomainday/2011/pddfaq/> accessed on 9th February 2025.

7 See, Seun Timi-Koleolu and Olawale Atanda 'Analysis Of Nigeria's National Artificial Intelligence Strategy' available at <https://www.mondaq.com/nigeria/new-technology/1507214/analysis-of-nigerias-national-artificial-intelligence-strategy> accessed on 20th February 2025.

iii. Accelerating AI Adoption and Sector Transformation

iv. Ensuring Responsible and Ethical AI Development

v. Developing a Robust AI Governance Framework

A key component of this strategy is the establishment of an AI Ethics Expert Group (AIEEG), an independent body tasked with overseeing the ethical use of AI and providing guidance on its responsible implementation. Additionally, the strategy introduces a standardized assessment tool to ensure that AI projects align with established ethical principles while also emphasizing the importance of a legal framework that protects human rights and privacy.[8] To foster innovation and investment, the NAIS emphasizes the availability of high-quality data for AI research and development, alongside holistic data governance standards that align with the Nigeria Data Protection Act (NDPA).[9] The strategy also highlights the need for public sector data availability, which can stimulate AI-driven innovation across critical sectors.

While this strategy sets a broad roadmap for AI adoption in Nigeria, it does not yet impose binding legal requirements on AI training or usage. This raises critical questions about how AI developers, regulators, and policymakers should approach the challenge of balancing innovation and compliance, particularly when it comes to the use of publicly available data in AI training. Some existing laws may help define what is permissible in AI training. The Copyright Act is crucial in determining the extent to which copyrighted materials can be used, while the reliance of generative AI models on vast amounts of personal data raises significant data protection concerns, making the Nigeria Data Protection Act (NDPA) and the Nigeria Data Protection Regulation (NDPR) highly relevant.

3.1 **AI Training and Copyright Law**.

One of the key legal areas affecting AI training is copyright law. With AI models generating massive data, many of which contain copyrighted works, developers and legal professionals are engaging with questions of infringement and liability. The copyright implications of data scraping for model training in generative AI tools have become contentious, leading to lawsuits and congressional hearings, especially as the

8 The establishment of an AIEEG is provided for under the fourth strategic pillar in the National Artificial Intelligence Strategy.

9 The need for holistic data governance standards in Artificial Intelligence development that conform with the NDPA is provided for under the third strategic pillar of the National Artificial Intelligence Strategy.

tools become commercialized and lucrative.[10] While there are no specific legislations regulating AI training in Nigeria, the Copyright Act 2022[11] could nonetheless serve as a useful guide by outlining clear parameters for the use of copyrighted materials in AI training and development. The Act grants copyright owners' exclusive rights over their works - rights that include reproducing, distributing, adapting, translating, making derivatives and communicating these works to the public.[12] Training AI models often involves data scraping, text and image analysis, or pattern recognition and as a result, these AI models may inadvertently perform actions covered by these exclusive rights, in processing copyrighted materials, especially when the outputs generated resemble or incorporate aspects of the protected content.

There are, however, certain exceptions that may seem to permit limited use of copyrighted materials without prior authorization. The doctrine of fair use has been relied upon by AI developers in arguing that their actions do not constitute an infringement. A popular instance would be the suit filed against OpenAI by New York Times accusing them of training their Large Language Models on material copyrighted by the Times.[13] OpenAI defends its methods by claiming that since AI models like GPT-4 are made to create new content by processing and synthesizing vast amounts of text, using data like articles from The New York Times and Daily News falls under fair use. OpenAI contends that neither express consent nor payment to the original creators is necessary for this process, which entails learning from a wide range of publicly accessible materials as the use is transformative.[14]

A recent ruling in *Thomson Reuters v. Ross Intelligence* [15]suggests that fair use may not always apply when AI systems systematically extract and repurpose copyrighted content. In this case, Ross Intelligence was found to have improperly used Westlaw's

---

10 See, Bruce D. Sokler, Alexander Hecht, Christian Tamotsu Fjeld and Raj Gambhir '(Un)fair Use? Copyrighted Works as AI Training Data — AI: The Washington Report' available at<https://www.mintz.com/insights-center/viewpoints/54731/2024-01-10-unfair-use-copyrighted-works-ai-training-data-ai> accessed on 9th February 2025.

11 Copyright Act, 2022.

12 Section 9, Copyright Act 2022.

13 See, Katherine Klosek, Marjory S. Blumenthal 'Training Generative AI Models on Copyrighted Works Is Fair Use' available at <https://www.arl.org/blog/training-generative-ai-models-on-copyrighted-works-is-fair-use/> accessed on 9th February 2025.

14 See, Kevin Dravon 'OpenAI Faces Backlash for Deleting Evidence in NY Times Copyright Case' available at <https://www.icasr.org/news/openai-faces-backlash-for-deleting-evidence-in-NY-times-copyright-case> accessed on 16th February 2025.

15 See, Matt Growcoot 'First Legal Ruling on AI, Copyright, and Training Data Goes the Way of Creators' available at <https://petapixel.com/2025/02/19/first-legal-ruling-on-ai-copyright-and-training-data-goes-the-way-of-creators/#:~:text=Thomson%20Reuters%20has%20won%20an%20early%20victory%20for,in%20its%20copyright%20infringement%20lawsuit%20against%20Ross%20Intelligence>. accessed on 19th February 2025.

curated legal materials to train its AI-powered research tool, with the court holding that such use was insufficiently transformative and directly harmed Westlaw's market. This decision raises questions about whether similar reasoning could apply to generative AI models trained on vast amounts of copyrighted text.

In Nigeria, the Copyright Act outlines the doctrine of fair dealing, which permits use of copyrighted material for specific purposes such as private study, non-commercial research, criticism, review, reporting of current events, and transformative applications.[16]

AI training for research and academic purposes should ordinarily find justification under this fair dealing provision, but commercial AI applications, such as generative AI models used in business are unlikely to be protected under fair dealing. This is because the determination of whether an AI model's training process would qualify as fair dealing depends on factors such as the purpose of use, the nature of the work, the portion used, and its effect on the market for the original work.[17]

Apart from these exceptions to the rights of a copyright holder which would most likely result in legal battle, AI developers seeking to utilize the creative works of authors for the process of AI training have the safer option of obtaining licenses or seeking assignments from these authors.[18] For AI training, this provision in the Copyright Act 2022 clarifies how copyright ownership and licensing function when using copyrighted materials. It establishes that copyright is a transferable property, meaning in this context, AI developers must obtain legal authorization through assignment or licensing before using protected works in training datasets. In situations where an AI developer intends to use such work exclusively, likely to establish a competitive edge over other developers, the licensing agreement must be reduced into writing.[19] Non-exclusive licences, however, do not require writing to take effect. It is important to highlight that copyright holders can license both existing and future works, but agreements cannot cover all future works of an author, preventing blanket claims over an individual's entire creative output.[20]

16 Section 20, Copyright Act 2022.
17 *Ibid.*
18 See section 30, Copyright Act 2022.
19 Section 30(3).
20 Section 30(10).

3.2 AI Training and Data Privacy

As a result of personal data often implicated in the process of training AI models, the regulations in different jurisdictions intended to protect privacy, ensure data security, and uphold fundamental rights are relevant.

In Nigeria, the Nigeria Data Protection Act, enacted in 2023 and the Nigeria Data Protection Regulation (NDPR) 2019 issued by the National Information Technology Development Agency (NITDA) provide the legal framework governing personal data processing.

3.2.1 NDPR/ NDPA Position

Certain provisions of the NDPR 2019 deal with ethical concerns that may arise in processing personal data generally. These provisions seem to be relevant in the protection of personal data use for Artificial Intelligence training.

Under Regulation 2.1 of the regulation, any personal data being processed must be collected and processed only for a clearly defined, legitimate purpose for which the data subject has provided explicit consent. This suggests that data incorporated into training datasets must be gathered with the individual's informed permission and may only be reused for narrowly defined purposes such as archiving, scientific or historical research, or statistical analysis in the public interest.[21] In addition, the data must be accurate, sufficient, and handled in a manner that respects human dignity, stored only as long as it is needed for the intended purpose, and rigorously secured against any foreseeable risks such as cyberattacks, physical damage, or unauthorized dissemination. Furthermore, anyone who handles or processes such personal data is required to exercise a high degree of care and is fully accountable for all actions taken with that data, ensuring that AI training processes are conducted responsibly and in strict compliance with these protective principles. The Nigeria Data Protection Act 2023 also establishes a framework that mirrors the NDPR's approach to processing personal data for AI training. The Act mandates in Section 24(b) that personal data must be collected and processed solely for clearly defined, legitimate purposes.[22] It also emphasizes that such data should only be processed based on explicit consent of the data subject concerned or a lawful basis, ensuring that any data incorporated into training datasets is gathered with informed permission and strictly used for narrowly defined

[21] Reg. 2.1, Nigeria Data Protection Regulation 2019.
[22] Section 24 (b) Nigerian Data Protection Act, 2023.

purposes.[23] This provision is analogous to Article 6 of the General Data Provision Regulation (GDPR) which also requires consent of a data subject before their personal data can be processed.

Another important provision for consideration in the Nigeria Data Protection Regulation (2019) is in Reg. 2.3 which requires that before any personal data is collected, the specific purpose for its collection must be made clear to the data subject, and that the consent obtained must be devoid of any fraud, coercion, or undue influence. Importantly, if personal data is to be transferred to any third party, the privacy policy must explicitly disclose this fact along with details of the intended use and the reasons for such a transfer. In Nigeria, personal data privacy has been receiving increased attention in recent times. and has been recognized as a fundamental right, with the courts affirming its protection under Section 37 of the 1999 Constitution,[24] which guarantees the right to private and family life, and have therefore been held to be actionable under the Fundamental Rights (Enforcement Procedure) Rules (FREP), as demonstrated in *Folashade Molehin v. UBA*.[25]

Recently, a class action suit was instituted against LinkedIn for allegedly disclosing customer information to third parties for the purpose of training artificial intelligence models without adequately informing users or obtaining their explicit consent. According to the allegations, LinkedIn's privacy policy failed to clearly disclose that customer data might be transferred and used for AI training.[26] Data controllers who may need to share personal data of data subjects must therefore seek the express consent of these data subjects before engaging in such conduct.

4 CONCLUSIONS

In conclusion, the dual challenges of privacy and copyright in AI training demand a balanced approach that protects individual rights and creative output while promoting innovation. Given the complexities that arise with this process, AI developers and companies must adopt a range of strategies. They should implement a "privacy by

---

23 Section 25 (1) (a) Nigerian Data Protection Act, 2023.

24 Section 37, Constitution of the Federal Republic of Nigeria 1999 (as amended)

25 See, Festus Oguns '2024: Review of Significant Decisions on Fundamental Rights Enforcement' available at <https://thenigerialawyer.com/2024-review-of-significant-decisions-on-fundamental-rights-enforcement/> accessed on 20th February 2025.

26 See, Jonathan Stempel 'Microsoft's LinkedIn sued for disclosing customer information to train AI models' <https://www.reuters.com/legal/microsofts-linkedin-sued-disclosing-customer-information-train-ai-models-2025-01-22/> accessed on 11th February 2025.

design" framework, ensuring that personal data is anonymized or pseudonymized wherever possible and that data processing practices are transparent and fully disclosed to data subjects.[27] Equally important is securing all necessary licenses for copyrighted content and establishing robust internal governance mechanisms to regularly audit and verify compliance with both privacy and intellectual property laws. Regulators and policymakers, meanwhile, must continue to refine and harmonize these legal frameworks to keep pace with rapid technological advances, ensuring that they provide clear guidance while not unduly hindering innovation.

For further information on this article and area of law,
Please contact **Chidumebi Nwosu** at
S. P. A. Ajibade & Co.,
Lagos by Telephone (+234 814 594 2221, +234,811 890 3060)
Fax (+234 806 437 3384)
Mobile (+234. 703 9141 316)
Email:cnwosu@spaajibade.com
www.spaajibade.com

[27] See, Graham Thompson 'Data Anonymization in AI: A Path Towards Ethical Machine Learning' available at <https://www.privacydynamics.io/post/data-anonymization-in-ai-a-path-towards-ethical-machine-learning/> accessed on 11th February 2025.