Copyright Infringement in Data Scraping During AI Model Training

Lexuan Tong

NFLS Yuhua International Academy， Nanjing，210012，China
Corresponding Author: Lexuan Tong, Email: SaraTong0923@outlook.com

Abstract
The rising pace of generative AI has brought copyright infringement relating to training data scraping to the very fore of debate in the legal and technological communities. This paper systematically examines the key questions of whether, and under what specific circumstances, the unauthorized scraping of copyrighted materials for AI model training constitutes copyright infringement. Through a literature review paradigm, this research examines two converse streams in academic positions: non-infringement and infringement, alongside key doctrines of the law upon which the case relies. In order to referee this conflict, a novel framework with a focus on the threefold core of copyright-creativity, authorship, and the extension of public knowledge is formed. It is within this paradigm that the above-mentioned study employs two major oppositions between ideas and expression as well as expressive use and non-expressive use in order to define the parameters of reasonable use in relation to AI across the data acquisition, storage, processing, and output generation pipeline. The core conclusion posits that data scraping for AI training is not inherently infringing. Liability exists if it replicates the essential original expression of a work and/or if it leads to harmful impacts market value, or exceeds the remit of a possible non-expressive function. The resulting balance attempts to correlate and measure the rights of creators with the mandates of purely technological advancement and knowledge-sharing, and can therefore be seen as providing useful criteria for dealing with current and future issues of a legal adaptation

Keywords
Copyright Infringement, AI Model Training, Data Scraping, Reasonable Use

Introduction

According to Rosenfeld et al. (2025), Thomson Reuters Enterprise Centre GmbH filed a lawsuit against the AI startup ROSS Intelligence in the United States in 2020. It centers on the critical question of whether using copyrighted legal head note to train a competing AI research tool constitutes transformative fair use or a market-harming infringement of the original expressive work. Similarly, copyright disputes have also arisen in the field of artistic creation. Three artists filed a copyright infringement lawsuit against Stability AI, Deviant Art, and Mid journey at the same time. The defendants were accused by the plaintiffs of direct copyright infringement in that they made use of artificial intelligence to scrape information and produce images that were similar to the original artwork of the artists (Xue & Zhang, 2025). All of these cases raise legal concerns regarding the

Citation: Lexuan Tong. (2026) Copyright Infringement in Data Scraping During AI Model Training The Journal of Young Researchers 1(3): e20260420

Received on March 6, 2026; Accepted on March 12, 2026; Published on April 20, 2026

unauthorized usage of the copyrighted content during training AI.

This dissertation attempts to address the questions of whether data scraping constitutes infringement and in what circumstances such an act amounts to infringement. The Generative AI system is unsupervised and semi-supervised with certain learning abilities and capabilities for production. This AI system acquires knowledge from very large amounts of data and generates new results that "have the same statistical characteristics as the data, but are not identical copies" (OECD, 2018; Hu, 2022).

For the discussion on the topic, this dissertation will be structured into three sequential parts. The first part will be the literature review. I will summary and analyze key academic viewpoints through clarifying the concepts. In the second part, I will elaborate on the conditions of AI copyright infringement by first discussing existing theories of copyright protection and then proposing the proper principles of copyright infringement in the AI era. In the final part, I will conclude that AI data scraping does not inherently constitute infringement and provide my evidence to prove that.

Literature review

The learning process of generative AI fundamentally relies on a technique known as data scraping, which involves identifying information sources, having programmers define search criteria, and then using software to interact with the target data source to extract large amounts of information that meet the criteria (Anina Ot, 2023). Kazmali & Sayar (2025) offer a clear definition of web scraping. It is a technique to automatically extract structured data from online resources. This method is used for gathering information in particular formats. These formats include Hyper Text Markup Language (HTML), Extensible Markup Language (XML), and JavaScript Object Notation (JSON). Web scraping can be done manually by the researcher, or software (Kazmali & Sayar, 2025).

Regarding the litigation on the infringement of data-scraping issues, there are academics and legal practitioners are essentially web scrapers and focus heavily on web scraping practice. The Perplexity AI case illustrates Perplexity's Answer Engine automatically surfs onto the sites of both Britannica and Merriam-Webster, which employs crawler software or other automated tools for targeted searching or systematic web scraping of content from these websites. Its core legal dispute is over whether the unauthorized scraping of copyrighted online content-such as encyclopedic text and dictionary definitions-for AI training purposes constitutes copyright infringement (Brittain, 2025).

In the existing literature, there are two opposing views on whether the data scraping and training of AI constitute copyright infringement. The first is referred to as the "non-infringement position," which holds that there is an essential difference between AI training replication of works and the replication and dissemination of original works. The second is termed the "infringement position," which argues that the core of the infringement dispute lies in procedural infringement.

Non-Infringement Position

The first position asserts that the act of copying works for AI training should not be viewed as equivalent to copying for reproduction or distribution of the original work. In literature, there are two kinds of arguments supporting this position.

Non-expressive Use and Market Non-competitiveness of AI Training

The first kind of argument is proposed by Wolfson (2023). He believes that using works as training data for AI operates in a market entirely distinct from that of the original works (Wolfson). In the original work market, the primary subject of transaction is expressive content. According to Cana (2023),

In his 1981 edition of Learning to Look, Joshua C. Taylor introduces the term expressive content to describe the creative combination of subject matter and form that draws from the viewer a response that is both unique to him and universal.

Consumers in this market focus their demand on expressive uses of works, such as appreciation, dissemination, or commercial display. By contrast, AI training does not target or exploit the expressive content of works. The utilization of a copyrighted work in this context does not involve sharing the work's original creative expression with the public; instead, it is limited to the extraction of practical, factual details or the underlying structural patterns embedded in the data (Sag, 2019). It means that AI primarily aims to extract non-expressive features from data. According to Amanda Ross, "Non-expressive use refers to uses that involve copying, but don't communicate the expressive aspects of the work to be read or otherwise enjoyed" (Ross, 2024). A clear example of non-expressive use in AI training is when a large language model processes the text of novels to learn statistical patterns of word sequences, without retaining or subsequently outputting any specific, protected sentences or plots from the original works. Thus, AI training with copyrighted works does not produce direct competition with the market for original works.

According to Sag (2019) the compatibility of this claim with scenarios where AI models memorize and reproduce the core content of original works lies in distinguishing between temporary technical memory in technical processes and communicative reproduction to the public. The key is judging the purpose of memorization rather than the act of memorization itself. In terms of the core logic of theoretical compatibility, non-expressive use does not absolutely exclude temporary technical memory of original works during AI training. For example, when an AI system temporarily caches text data to extract language grammar or data statistical rules, legal analysts note that this technical memory is only necessary for the training process and does not convert the original expression of the work into content that the AI disseminates to the public, so this use still aligns with the core of the non-expressive use principle. Sag also claims that AI training does not convey the original works' expression but rather uses the data to learn patterns, making it a form of non-expressive reproduction that should not be considered infringement. For example, in the Authors Guild's copyright infringement lawsuit against HathiTrust Digital Library (Law, 2014), the plaintiff alleged that HathiTrust's scanning of millions of copyrighted books into digital formats infringed their reproduction and communication rights. However, the court ruled this constituted "non-expressive use" under fair use, as the full-text search function merely converted works into researchable data without exploiting or disseminating their expressive content. The principle of fair use, proposed by American Justice Joseph Story in *Folsom v. Marsh* (1841), is a core principle that prioritizes the inherent property interests of authors in the fruits of their labor over the utilitarian goal of progress in the arts (Rutkowski & Won, 2025). Buick (2025) also supports this view in his research. He states that when training artificial intelligence, courts should not categorize temporary reproductions of the same as a copyright violation since the reproductions is not a direct copy of the original work.

The Purposes of Copyright Legislation and Public Interest Orientation

The second kind of argument is based on the conception of the prime role of intellectual property law. Quang (2021) contends that the use of copyrighted content by the developers who operate the machine learning systems to train them is consistent with the ultimate purpose of the intellectual property law, which is to maximize the amount of intellectual good that can be enjoyed by the people. According to Quang, the use of copyrighted materials to come up with such systems facilitates technological advancements and adds to the overall development of the society. In addition, Microsoft has stated in court proceedings that obliging its use to seek formal permission each and every time it uses a piece of copyrighted content will impede advancement in machine learning technology. This will disproportionately affect smaller developers in this region (Oppedal, 2024). This view is based on normative approach. In this area of research, one should be primarily concerned with the progressive growth of technology and the innovation practices, the

value system of which does not rely on the strict adherence to the rights of the holders of copyright. According to Kumar (2024), innovation is a constantly expanding source of societal development, as well as a force of economic growth. According to him the protection clauses that are offered in the name of the intellectual property right (IPR) regimes are a necessary stimulus in the technological growth and development. This kind of environment will enable individuals and companies to derive the benefits of intellectual work. Kumar further adds that it is only under such intellectual property protection that innovation can be the driving force behind economic and social advancement and growth and such a factor encourages the exchange of knowledge on a global scale.

This position argues aligns with the constitutional underpinnings of U.S. copyright law， illustrating that using copyrighted works to train AI does not constitute direct infringement. The Copyright Clause of the U.S. Constitution prioritizes promoting the progress of science and useful arts. AI models drive innovative development by extracting data patterns. This is essentially the reuse and expansion of public knowledge, which is highly consistent with the constitutional purpose of the copyright system and should not be simply classified as infringement (Congress of the United States, 1789).

Infringement Position

Despite the arguments for non-infringement, scholars who hold the "infringement position" argue that the core point of contention in relevant copyright cases lies in process-based infringement: AI companies use copyrighted works for the development and training of their AI models without obtaining the prior authorization of copyright holders. In existing academic literature, there are three kinds of arguments supporting this position.

Idea-Expression Distinction: A Core Principle for AI Copyright Boundaries

The first kind of argument, the idea-expression distinction, proposed by TRIPS, is the fundamental principle of Intellectual property, especially in copyright (WTO, 1994). According to Pranav (2024), this means that there has to be a clear demarcation line in which remain accessible to the public knowledge and understanding. Under this distinction, if an AI system persistently saves the full content of original works or holds the core original expression of the work in order to identify non-core ideas from the AI's the use of the works exceeds the boundary of legally permissible use. This action of storage may violate the reproduction rights afforded to copyright holders. One famous example is that of *Facebook, Inc. v. Power Ventures, Inc.* (2019) in which the court held that scraping the entire user profile pages could constitute infringement. The reasoning of this ruling was that Facebook had the rights to protect the copyright of the overall structure and the expressive features of the pages, although the content created by the user of the center pages was not covered by such protection (Neudata, 2024).

Balancing Copyright Enforcement and AI Technological Innovation

The second kind of argument revolves around the conflict between implementing the copyright law and the advancement of technology to make novel AI. Li et al. (2024) found that there is some need to protect the rights of authors. The overprotective copyright regulations might affect two valuable aspects of AI development. Firstly, these initiatives could hamper innovation capabilities in the industry. Secondly, they might pose a barrier for small-scale developers in their access to training necessities for their models. It argues that a good balance has to be achieved. It has to solve two basic elements. On one hand, it is necessary for it to take into account the rights of copyright holders. On the other hand, they also ought to factor into consideration the wider benefits that come with the development of AI.

Copyright Exclusive Rights

The third line of argument is on the essence itself. This right is regarded as the right or privilege that can only be used by the person who is granted this particular privilege to (Staff, 2013). According to Patterson (1993), the constitutional basis of copyright is to grant authors a limited monopoly on publication and sale of their

literary works. On this premise, it principally depends upon the regulation of reproduction acts. The organized scraping and the reproduction of work on machines such as AI without authorization erodes the integrity of the exclusive rights prior to the works falling into the public domain, before they enter the public domain (Patterson). Ginsbur (2017) puts particular stress on the importance of ensuring, especially in the age of the Internet, that the right of authors to control the use of works is key to realizing copyright's public function of advancing knowledge. Accordingly, despite the transformative nature of AI works, any unauthorized copies of copyrighted works during the training process, is considered an infringement on the author's exclusive rights (Ginsburg).

Discussion

The above review shows an impasse in the scholarly discourse. One position justifies AI data scraping as a non-expressive, driven activity the final aim of the copyright, while the other considers it as an infringing process that affects creators' exclusive rights. This ongoing disagreement implies that there are underlying conceptual issues. The following discussion will continue beyond the description conflict of positions in developing a framework of analysis in adjudicating claims of infringement.

The Core Interests Of Copyright Protection

According to WIPO, copyright is a legal concept that defines the rights that creators of intellectual property have with respect to literary and artistic work. There are two kinds of rights that copyright laws recognize. The first is economic rights. Such rights entitle the copyright holder to seek financial gain through the use of others in their work. The second is moral rights which relate to the author's non-juridical interests (WIPO, 2024). The above definition has clearly stated the framework of copyright as an instrument of law, wherein the use of works is governed by economic rights, moral rights. However, an inquiry into the internal values these rights seek to protect reveals that the fundamental concerns of the copyright system can be distilled into three aspects: safeguarding authorship to preserve the moral character of creators, encouraging innovation to incessantly enrich the product of social culture, and ultimately, it is achieved through the equalization of rights, promotion of accumulation and dissemination of public knowledge. Therefore, creativity, authorship, and public knowledge collectively embody the threefold basis of copyright protection.

Creativity as the Threshold and Core Purpose of Copyright Protection

Creativity serves as the primary threshold for copyright ability. "Creativity is defined as the tendency to generate or recognize ideas, alternatives, or possibilities that may be useful in solving problems, communicating with others, and entertaining ourselves and others" (Franken, 1994). Creativity is crucial in copyright law. First, it is the core basis for setting the threshold of copyright protection. Copyright law does not protect ideas themselves, but only their original expressions. The ideas, alternatives, or possibilities embodied by creativity are the core components of originality. Only works reflecting the creator's independent thinking and personalized expression choices are eligible for protection. On the contrary, facts, general methods, or mechanical compilations lacking creativity cannot obtain protection. What's more, creativity is the key carrier for achieving the incentive purpose of copyright legislation. Copyright protection finds its direct constitutional footing in U.S. law. The framers of the Constitution held the view that granting authors exclusive rights to their original works for a restricted duration would serve to advance the development of science and practical arts (Copyright Alliance, 2020). Works created with creativity, which have both practical and spiritual value, can bring economic rewards to creators through exclusive copyright rights.

Authorship: The Secondary Threshold and Legal Foundation of Copyright Protection

Authorship serves as the secondary threshold for copyright ability. It is not a right that grants legal protection to all acts of selecting, coordinating, and arranging elements, but only to those carried out in a specific manner and capable of

generating psychological effects (Buccafusco, 2016). I believe authorship's importance is primarily reflected in the following two key dimensions. On one hand, it is grounded in ownership and value added through labor, being the partial rights in their work prior to registration with the Copyright Office. Without the recognition of authorship, subsequent actions such as rights transfer and licensing would lack a legitimate source. In the meanwhile, legal recognition of authorship is functionally equivalent to invoking the attribute of respecting the intellectual labor value of creators. This connects invisible intellectual labor to visible rights and benefits and a justifiable reason for the creators to seek economic returns. In essence, this amounts to a legal affirmation of a basic natural right. As the U.S. Supreme Court held in The Antelope (1825), every man has a natural right to the fruits of his own labor, is generally admitted; and that no other person can rightfully deprive him of those fruits, and appropriate them against his will, seems to be the necessary result of this admission. (23 U.S. 66) Another point worth discussing is how the law illustrates the highly important role of originality with regards to an individual's work and creative effort. Authorship has two important functions to play. To begin with, it serves as the foundation of protection of original work by a creator. Second, the act serves as an incentive in the wider innovation and development in the society. It encourages even more creators to practice the work of imagination and innovativeness in creating certain legal standards and positive incentives. It also helps in creating plural of original things and new discoveries. This in turn speeds up translation and transmission of such novel work which eventually results in the further development of the cultural and technological aspects of society.

Public Knowledge Expansion: The Final Threshold of Copyright Protection

The final threshold of copyright capability is the public knowledge expansion. It also involves expansion, enriching, and utilizing the knowledge base that is accessible to the collectivity. In my opinion, the growth of people knowledge is the mechanism of social culture prosperity. It has implications in the three following aspects. First, it gives energy to the whole intellectual resources deposits in the society. It enables interpersonal boundaries to be broken by knowledge in other fields. Second, it has minimized the obstacle to knowledge acquisition and distribution. It also questions the monopoly of knowledge that some groups of people possess and facilitates the acquisition of knowledge across community lines. Finally, it also sparks the life of knowledge re-innovation. The reference, integration, and breakthrough of the existing knowledge are a rise of new theories and technologies. The increase of the system of public knowledge gives the innovators more different materials and views, creating a vicious circle of creation, sharing, re-creation.

Balancing Copyright's Core Elements: The Role of Authorship Restriction

In the framework of copyright law, the three core elements, authorship, expansion of public knowledge, and creativity, form a mutually reinforcing whole. Among them, the demarcation of authorship boundaries is the key to maintaining this balanced relationship. Because when trying to protect the authorship while stimulating creativity, conflicts caused by the exclusivity of private rights are unavoidable and will contradict the sharing of public knowledge. Under U.S. copyright law, "exclusivity" refers to the owner's sole right to perform and authorize others to perform six key acts (17 U.S.C. § 106, as cited in GovInfo, 2024). As elaborated earlier, authorship is not an unrestricted right, but a qualified entitlement. Such restrictions are not designed to undermine creators' legitimate interests, but to guard against the over expansion of exclusive rights. It can effectively prevent them from becoming barriers that hinder the expansion of public knowledge. In summary, by means of avoiding the monopolization of rights, preserving room for innovation, and facilitating the circulation of knowledge, it builds an institutional framework by being characterized by encouraging creation, knowledge sharing, and re-innovation. Finally, the sustainable development of creativity is accomplished.

Creator's Exclusive Rights

But, "infringement position" academics propose that such protection should: preserve the sole author's rights upon the work.

Historical Evolution of Copyright: Balancing Rights Protection and Innovation

From a historical viewpoint, the development of copyright protection from prohibiting all copying to authorizing limited copying can be said copying to fair use. This is actually a search on the better balance between right protection and promotion of innovations. This legislative intent would be destroyed by giving absolute prominence to the exercise of the right of authorship. It would bring a change in the law of copyright being a source of innovation to an instrument in the creation of monopolies. This is because the concept of the copyright monopoly lacks monopolies in the rights to their work which should not be encouraged. This does not coincide with the two key goals of the copyright laws to encourage invention and facilitating the accessibility of public knowledge to grow and disseminate. Exclusive rights are the ones that are owned by the creators of copyrighted works and are based upon the principle of the right of free domination of their works by all the means. With such a condition, even the non-commercial transformative uses that extract solely abstract language modes will be subject to stringent restrictions. Too much expansion of these rights may give birth to unsurmountable obstructions for sharing knowledge, but also innovations. For example, the Berne Convention stipulates that the copyright protection period lasts "the author's lifetime plus 50 years after death," after which the work enters the public domain. This norm itself embodies a system design that aspires to transfer individual rights to the public interest. It also corresponds to the goal of spreading sociological progress.

Rejecting Exclusive Rights: A Rational Safeguard for Authorship Essence

Second, the denial of exclusive rights does not mean the denial of authorship. On the contrary, it seems to represent a more rational way to protect the nature and worth of authorship. Authorship is essentially the starting point for all copyright determinations. This is because not only is authorship crucial for deciding that protects the intellectual work of authors. Going beyond this responsibility, its significance cannot be overemphasized within original work. If authorship were disregarded, dreadful results would follow. The authors would be denied the rightful economic benefits. They would also lose moral rights tied to their works, such as the right of attribution. Such deprivation would severely dampen their creative motivation. Opposing exclusive rights is not intended to negate authorship, but to prevent it from being distorted into a tool for rights abuse.

The Definition and Practice of Creators' Core Essential Control Rights

Last but not least, authorship should vest creators with core and essential control rights. An excessively minimal level of control would fail to effectively protect creators' fundamental interests, while exclusive rights would stifle the vitality of knowledge circulation and re-innovation. The core essential control refers to the exclusive dominion that creators hold over the core commercial and moral interests of their works. The core essential control specifically encompasses the following three aspects. First, it should enable creators to obtain economic returns from the commercial reproduction, distribution, and adaptation of their works. Second, it should confer upon authors the right of attribution. Third, it needs to endow authors with the right to protect the integrity of their works and prevent them from being distorted or mutilated. In the *Authors Guild v. OpenAI* (2025), the judge dismissed the plaintiff's claims on the core reason that OpenAI's AI training only extracted abstract language patterns, did not replicate the core expression of the work, and did not harm the market for the original work. This means that functional use does not constitute infringement if it does not infringe upon the creator's core rights such as the right to market revenue. In this way, such core essential control achieves a precise balance. It not only safeguards the legal rights and interests of creators but also provides space for the expansion of public knowledge and technological innovation.

(Authors Guild v. OpenAI, S.D.N.Y. 2025, Doc. 782 at 27)

The Principle of Reasonable Use and Its Application in the AI Era

Since authorship should not be exclusive, reasonable use is the key institutional design for balancing authorship, expansion of public knowledge, and creativity. The following discussion will focus on the core values, scope of application, and criteria for determining reasonable use. Through analysis of the technical aspects of data scraping on artificial intelligence, the boundaries relating to conditions and infringement would be explained in this context, offering an operational framework that may be used as a guideline in AI era.

On the Core Value and Definition of Reasonable Use

The essence of the reasonable use principle is based on its function as a fundamental tool to balance copyright holders' exclusive rights, the public interest in social knowledge sharing, and the demand for technological innovation. It is the key link to achieve the dual purposes of copyright legislation: promoting creation, encouraging the dissemination of knowledge. In essence, reasonable use refers to the act that the law permits to the public to use protected works without the copyright holder's permission or payment on certain conditions. What fundamentally characterizes fair use is that it shall not impair the legitimate right and interest of the copyrighter or disturb the normal publication and spreading of a work.

The Core Boundaries and Criteria of Reasonable Use

The scope of application of this principle must adhere to two core boundaries. First, it must not violate the basic content of authorship. In other words, the user must not seize the creator's status as the original producer of the work. It also must not harm the moral rights tied to authorship, such as the right of attribution and the right to protect the integrity of the work. Second, it must respect the copyright holder's core essential control rights. More specifically, it must not violate the exclusive rights of the creators over their works, such as the rights to commercial reproduction, distribution, adaptation, and the right to obtain economic benefits from these acts. It is also not supposed to cause substantial substitution or damage to the market value of the original work.

Judging Unreasonable Use in AI Scenarios

The core idea-expression distinction and the expressive-non-expressive use distinction shall be considered to examine whether a use constitutes unreasonable use and thus infringes on the copyright holder's exclusive rights. Firstly, as confirmed by the idea-expression distinction, copyright law protects only the original expression of a work. If an AI data scraping and training act goes beyond the extraction of ideas, facts, or general patterns, but rather copies or uses the core original expression of the work, it will constitute infringement of exclusive rights. Secondly, according to the distinction between expressive and non-expressive use, if the core purpose of the use is to convey the expressive content of the work and spread the work's aesthetic or informational value to the public, it is an expressive use. On the contrary, if the use only extracts the non-expressive features of the work, for instance, linguistic grammar rules and data statistical patterns during AI training-and does not disseminate the original expression of the work to the public, it is a non-expressive use. This complies with the core essence of reasonable use and does not constitute infringement.

Determining Copyright Infringement of AI Training Data Scraping

To determine whether data scraping during AI model training constitutes copyright infringement, two levels of boundaries are needed to be clarified: general identification logic and specific case classification.

General Identification Conditions

From the perspective of general identification conditions, three elements must be met simultaneously for an act to be deemed infringement.

First, the object of use must be the original expression protected by copyright law,

excluding ideas, facts, general materials, and other public domain content.

Second, the AI data scraping must involve acts controlled by exclusive rights, such as reproduction, distribution, and adaptation.

Third, the use does not fall into the scope of rights restrictions, such as reasonable use or statutory license, and it causes substantial damage to the copyright holder's core essential control rights or the market value of the work.

Specific Infringement Scenarios

In terms of specific infringement scenarios, they can be divided into three categories.

First, the data scraping act copies the core expression of the work. For example, an AI system fully stores or copies large batches of original content of works during data scraping—such as original annotations of legal cases, text paragraphs of literary works, and high-definition images of artistic works. These directly copied contents are then used to train models to generate substitute content highly similar to the original works, directly replacing the market function of the original works.

Second, the data scraping act damages the market value of the work. For example, an AI company scrapes content from paid academic paper databases to develop free alternative literature search tools, or grabs works from paid image libraries to generate commercial images. This leads to a decline in market demand for the original works and directly infringes on the copyright holder's right to economic benefits.

Third, the data scraping act exceeds the boundary of non-expressive use. For example, during AI training, the system not only extracts data patterns but also uses technical means to memorize and reproduce the unique expressive details of the work—such as a specific writer's language style or a specific designer's iconic elements—and substantially reproduces these expressions in the output results. This enables the public to perceive the original expression of the work through AI-generated content, thus constituting an unreasonable infringement of exclusive rights.

Specific Behaviors and Potential Copyright Infringement Risks

The four basic steps of training a model of AI—data procurement, data storage, processing, and the production of outputs—can be assigned behaviors and risks involving copyright infringement.

1.Data Acquisition stage: This is where AI obtains elements of artistic works from data resources using means such as web crawlers. If this process is accompanied by procedural violations, such as disobeying authorization systems, but lacks the intent to obtain abstract rules, then it will lack the necessary elements required under non-expressive expression. This kind of act will infringe on the copyright holder's right of communication through information networks. It will also undermine the holder's legitimate control over work dissemination and the realization of labor value.

2.The data storage stage is the process where AI fixes and retains the acquired work data in electronic form. If the storage is non-temporary and not technically necessary, it does not involve data transformation. It only facilitates subsequent infringing activities. This kind of act will constitute an infringement of the copyright holder's right of reproduction. It will also improperly appropriate the labor value of creators.

3.The data processing stage is the process where AI analyzes stored data and extracts features. If the processing lacks transformativeness, it fails to make the leap from the expression level to the rule level. It only copies or slightly modifies the core expression of the original work. Alternatively, if it substantially uses the parts reflecting the author's labor, skills, and originality, and causes material harm to the derivative market of the original work (such as direct market competition), it will infringe on the copyright holder's right to remuneration.

4.The output generation stage is the process where AI generates new content based on processed data. If the generated content is substantially similar to the original work and

achieves no value-level transformation, it replaces the market circulation of the original work. It may even damage the author's reputation or the integrity of the work. This kind of act will infringe on the copyright holder's right of adaptation. It will also weaken the incentives for original creation.

Conclusion

This research systematically tackles the issue of copyright infringement lawsuits involving data scraping during the AI model-training process, answering the two fundamental questions of whether data scraping constitutes infringement and under what circumstances such an act amounts to infringement. By analysing the two different approaches that have been adopted by scholars on the topic, that is, the "non-infringement view" and "infringement view," the dissertation has created a well-balanced framework for analysis that takes into account the three-fold core interest in the field of copyright protection relating to creativity, authorship, and the expansion of public knowledge. The study ultimately concludes that data scraping itself is not inherently infringing. This is because the legal status of scraping will depend on whether the activity is legal. For example, reproduces the essential original expression embodied in a work that is protected by copyright will harm the market value of the copyrighted work.

By means of the idea-expression distinction, as well as the distinction between use as the logical pillars of this study, this investigation explains the extent of the principle of fair use that is applicable to within the context of AI and establishes criteria for infringement at all four levels of data acquisition, storage, processing, and output generation. This framework not only provides a basis for protecting creators' core rights—particularly their control over the core expression and market interests of their works—but also claims a necessary institutional space for AI technological innovation and the accumulation of public knowledge.

However, this study also has certain limitations. The dissertation depends on mainstream theories on the issue and mainstream scholarly discussions. Despite establishing a vast system for establishing infringement, I failed to integrate the divergent perspectives on the topic held by either scholarly or jurisprudence opinions. The developing and dynamic character of this branch of jurisprudence remains underutilised within the context of the study.

Copyright issues in the context of artificial intelligence have remained an unexplored field in the existing legal system. The worth of this study lies in the fact that it presents, phase by phase, a specific case-related study of each infringement case. It trys to offer a more flexible and futuristic approach to lawmakers and jurists in the future. It is the malleable nature of the law, in terms of its ability to suit the times, that makes it alive. Ultimately, interaction of law and technology is promoted in a mutually driving manner.

Conflict of Interests: the author has claimed that no conflict of interests exists.

References

1. Anina Ot. (2023, September 11). What is Data Scraping? Definition & How to Use it. Datamation. https://www.datamation.com/big-data/data-scraping/#how-data-scraping-works
2. Brittain, B. (2025, September 11). Encyclopedia Britannica sues Perplexity over AI "answer engine." Reuters. https://www.reuters.com/legal/litigation/encyclopedia-britannica-sues-perplexity-over-ai-answer-engine-2025-09-11/
3. Buccafusco, C. (2016). A Theory of Copyright Authorship. Virginia Law Review, 102(5), 1229–1295. https://scholarship.law.duke.edu/faculty_scholarship/4138/
4. Buick, A. (2024). Copyright and AI training data—transparency to the rescue? Journal of Intellectual Property Law & Practice, 20(3), 182–192. https://doi.org/10.1093/jiplp/jpae102
5. Cana. (2023, March 8). Cana Academy. Cana Academy.

https://www.canaacademy.org/blog/seeing-amp-moving-with-expressive-content
6. Congress of the United States. (1789). U.S. Constitution - Article I | Resources | Constitution Annotated | Congress.gov | Library of Congress. Congress.gov. https://constitution.congress.gov/constitution/article-1/#article-1-section-8-clause-8
7. Copyright Alliance. (2020, November 4). What Is The Purpose of Copyright Law. Copyright Alliance. https://copyrightalliance.org/education/copyright-law-explained/copyright-basics/purpose-of-copyright/
8. Counter Listener. (2014, June 10). Search Results for Courts: All › Query: Authors Guild v. HathiTrust › Published: True — 48 Results — CourtListener.com. CourtListener. https://www.courtlistener.com/?q=Authors+Guild+v.+HathiTrust
9. Franken, R. E. (1993). Human motivation (3rd ed., p. 396). Brooks/Cole.
10. Ginsburg, J. C. (2017). "The Exclusive Right to Their Writings": Copyright and Control in the Digital Age. University of Maine School of Law Digital Commons. https://digitalcommons.mainelaw.maine.edu/mlr/vol54/iss2/3/
11. GovInfo. (2024). United States Code, 2024 Edition, Title 17 - COPYRIGHTS. Govinfo.gov. https://www.govinfo.gov/app/details/USCODE-2024-title17/USCODE-2024-title17-chap1-sec106
12. JUSTIA. (1841). Folsom v. Marsh (D. Mass. 1841). Justia Law. https://law.justia.com/cases/federal/district-courts/massachusetts/madce/9fcas342/4104271/220/no-4.html
13. JUSTIA. (2023). Authors Guild et al v. OpenAI Inc. et al, No. 1:2023cv08292 - Document 782 (S.D.N.Y. 2025). Justia Law. https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2023cv08292/606655/782/
14. Kazmali, A. S., & Sayar, A. (2025). Web Scraping: Legal and Ethical Considerations in General and Local Context - A Review. Procedia Computer Science, 259, 1563–1572. https://doi.org/10.1016/j.procs.2025.04.111
15. Kumar, P. (2024). Intellectual Property Rights (IPR): Nurturing Creativity, Fostering Innovation. Ldealistic Journal of Advanced Research in Progressive Spectrums (IJARPS) EISSN– 2583-6986, 3(02), 32–38. https://journal.ijarps.org/index.php/IJARPS/article/view/321
16. OECD. (2018). Generative AI – The issues . Oecd.ai. https://oecd.ai/en/generative-ai-issues-overview
17. Oppedal, N. (2024). "Balancing Innovation and Copyrights: The Legal Framework for AI Training in the European Union."
18. Patterson, L. R. (1993). Copyright and "the Exclusive Right" of Authors. Digital Commons @ University of Georgia School of Law. https://digitalcommons.law.uga.edu/fac_artchop/343/
19. Pranav, D. (2024, April 25). Idea-Expression Dichotomy in Copyright: Judicial Rulings and Merger Doctrine. Khurana and Khurana. https://www.khuranaandkhurana.com/2024/04/25/idea-expression-dichotomy-in-copyright-judicial-rulings-and-merger-doctrine/
20. Quang, J. (2021). Does Training AI Violate Copyright Law? Berkeley Technology Law Journal, 36(4). https://doi.org/10.15779/Z38XW47X3K
21. Rosenfeld, J., Wood, S., Zoffer, H., McNeal, S. L., & Savage, C. W. (2025). Thomson Reuters v. Ross Intelligence: Copyright, Fair Use, and AI (Round One) | Davis Wright Tremaine. Dwt.com. https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2025/02/reuters-ross-court-ruling-ai-copyright-fair-use
22. Ross, A. (2024, March 2). LibGuides: Copyright Guide: Subscribe to News. Tulsacc.edu. https://guides.library.tulsacc.edu/copyright/News/fair-use-week-2024-day-four-with-guest-expert-dave-hansen-why-fair-use-supports-non-e

23. Rutkowski, J. D., & Won, K. K. (2025, July 16). Two Northern District of California Judges Find Some Copying to Train Generative AI Is Fair Use. Mintz.com. https://www.mintz.com/insights-center/viewpoints/2231/2025-07-16-two-northern-district-california-judges-find-some?utm_source=chatgpt.com
24. Sag, M. (2019). The New Legal Landscape for Text Mining and Machine Learning. Emory Law Scholarly Commons. https://scholarlycommons.law.emory.edu/faculty-articles/28/
25. Staff, T. (2013, March 28). EXCLUSIVE RIGHT. The Law Dictionary. https://thelawdictionary.org/exclusive-right/
26. U.S. Supreme Court. (1825). The Antelope, 23 U.S. 66 (1825). Justia Law. https://supreme.justia.com/cases/federal/us/23/66/
27. United States Court Of Appeals For The Ninth Circuit. (2019, January 18). FACEBOOK, INC. V. POWER VENTURES, INC., No. 17-16161 (9th Cir. 2019). Justia Law. https://law.justia.com/cases/federal/appellate-courts/ca9/17-16161/17-16161-2019-01-18.html
28. WIPO. (2023). Berne Convention for the Protection of Literary and Artistic Works. Treaties. https://www.wipo.int/en/web/treaties/ip/berne/index
29. WIPO. (2024). Copyright. Copyright. https://www.wipo.int/en/web/copyright
30. Wolfson, S. (2023, February 17). Fair Use: Training Generative AI. Creative Commons. https://creativecommons.org/2023/02/17/fair-use-training-generative-ai/
31. WTO. (1994, April 15). WIPO Lex. Www.wipo.int. https://www.wipo.int/wipolex/en/treaties/details/231
32. Xue, J., & Zhang, K. (2025). Copyright infringement in content generated by AI: an empirical analysis based on typical cases of China and the United States. Applied Mathematics and Nonlinear Sciences, 10(1). https://doi.org/10.2478/amns-2025-0813