Page: 691-707 *URL: www.lijdlr.com*

LAWFOYER INTERNATIONAL JOURNAL OF DOCTRINAL LEGAL RESEARCH

ISSN: 2583-7753

LawFoyer International Journal of Doctrinal Legal Research

[ISSN: 2583-7753]

Volume 4 | Issue 3

2026

DOI: https://doi.org/10.70183/lijdlr.2026.v04.290

THE DOCTRINAL OBSOLESCENCE OF TRADITIONAL EVIDENTIARY STANDARDS

Kushal Manish Jain[1]

I. ABSTRACT

Generative Artificial Intelligence (AI) in the form of Large Language Models (LLM) and latent diffusion image generators, has caused an intricate structural crisis in international copyright law. In particular, this crisis concerns with the unauthorized ingestion, reproduction and synthesis of protected works for machine learning training. The first challenge, evidentiary burden of proof, is a formidable obstacle in enforcing the rights of AI developers with respect to procedural IP considerations. In procedural IP considerations, the evidentiary burden of proof is a monumental and often insurmountable hurdle for enforcing IP rights of an AI developer. In essence, AI systems are algorithmic "black boxes," and it will be very hard for plaintiffs to establish that a specific creative piece was actually used in the training process, particularly if the output of the AI system is not visually or textually similar to the input. In this paper, the authors perform a comprehensive doctrinal examination of the evidentiary issues in modern AI copyright litigation. The research examines the principles of digital forensics, which are necessarily new and sophisticated, and explores their application with respect to traditional legal concepts like "access" and "substantial similarity," which are ultimately inadequate for the architecture of neural networks. The paper proposes a theory that the plaintiffs are forced to increasingly and exclusively depend upon sophisticated digital forensic techniques, such as metadata analysis, dataset tracing, hash matching and memorization extraction techniques to meet the evidentiary requirements of the courts without statutory requirements mandating transparency of the data sets. In conclusion, the research calls for the formal inclusion of standardised digital forensic procedures in IP litigation and suggests that there should be a legislative change in the direction of compulsory auditing of datasets to facilitate fair enforcement of copyright in the digital era.

[1] Student at National Forensic Sciences University (India). Email: kushaljain0506@gmail.com

 (ISSN: 2583-7753)

II. KEYWORDS

Generative AI; Copyright Infringement; Digital Forensics; Evidentiary Burden; Training Data Transparency.

III. INTRODUCTION AND RESEARCH PROBLEM

The history of copyright law, from the Statute of Anne (1710) to the Berne Convention for the Protection of Literary and Artistic Works (1886) and the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), has consistently reflected an attempt to balance competing public and private interests in creative expression.[2]Firstly the law is intended to confer a monopoly for a finite period of time on the creator of an original expression, to encourage cultural and scientific advancement. Secondly, it makes sure that these works over time become public domain and thus contributes to the collective knowledge of society. The emphasis of this legal structure is that "copying" is either a physical or a direct mechanical process. Copyright infringement claims rely on two essential elements: that the plaintiff owns a valid copyright and that the plaintiff's copyrighted work was copied without his permission from original constituent elements. In the past, copying has been circumstantially proved by showing that the defendant had "access" to the plaintiff's work and both works share a high degree of similarity.

The emergence of generative AI models like OpenAIs GPT architectures, Anthropics Claude and Stable Diffusion models from Stability AI, however, has thrown this longstanding evidentiary structure into chaos. The technologies that govern their development are based on a process called "machine learning" that involves feeding huge amounts of data, much of which is gathered without discrimination from the open Internet, shadow libraries, and proprietary databases. If you are using an AI company to train a model using billions of data points, the works themselves are not copied in a

[2] Berne Convention for the Protection of Literary and Artistic Works, Sept. 9, 1886, as revised; Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994.

traditional way and in an easily identifiable format. Instead, they are tokenized and embedded as statistical weights, biases and algorithmic parameters of a neural network.

The main research problem stated in this paper exists in this noticeable void of evidence. How can the copyright owner reasonably and legally demonstrate that a particular work was ingested and used by the AI developer if the training set is highly biased as a trade secret, the architectural process is black-boxed, and the art itself is not clearly or substantially similar to the original work? Digital forensics and intellectual property litigation are no longer theoretical, but a necessity in the procedures. If the system of substantive rights of the authors is not handled by the forensics, the rights are practically unenforceable against the well-capitalized technology companies.

A. Research Objectives

The principle goals of this doctrinal and interdisciplinary research are as follows:

1. First, it critically examines the doctrinal shortcomings of traditional evidentiary standards, such as the "access" and "substantial similarity" paradigms, in the context of the computational practices of machine learning and generative AI training methods.
2. Second, to broadly assess the growing importance of e-discovery data analysis as a tool to fill the evidentiary deficit in modern IP litigation in the context of data extraction attacks, preserving metadata, and conducting an API audit.
3. Third, to explore the balancing of a plaintiff's civil discovery rights in infringement cases and a defendant's rights to preserve proprietary training data under the trade secret doctrine.
4. Fourth, to suggest practical statutory, procedural, and technological changes that would require algorithmic and dataset transparency, which would help

to ensure the equitable administration of justice and uphold the basic aims of copyright.

B. Research Questions

In order to meet the above goals, the following basic questions will be addressed in this work:

1. How do opaque black-box architectures of generative neural networks pose a specific challenge to the plaintiff's burden in copyright infringement litigation?
2. How can prove digital forensic methods and technical mechanisms be legally accepted as evidence to support allegations that specific copyrighted materials were added to an unauthorized AI training dataset?
3. How are courts today resolving evidentiary conflicts between copyright plaintiffs and the trade-secret or corporate-confidentiality defenses raised by artificial intelligence developers?
4. Can existing international and national procedural rules and practices provide intellectual property tribunals and courts with the necessary resources and tools to assess, interpret, and decide highly voluminous computational and forensic evidence?

This paper rests on the following key hypotheses based on initial doctrinal observations and the state of the technology law landscape:

1. First, the traditional doctrine of proof for copyright infringement (access plus substantial similarity) is outdated and of poor design for machine learning algorithms that rely on non-literal copying mechanisms.
2. Second, digital data trails which have been successfully and inevitably extracted by courts in the past will become a key element of any successful prosecution and adjudication of AI-related copyright claims, in place of the traditional, visual, or textual comparison of outputs.

3. Third, as long as there are no statutory requirements for transparency in data or a presumption of ingestion for broadly disseminated digital works, AI producers will be able to effectively deploy a trade secret defense as evidence, leaving many human authors in a state of widespread disenfranchisement without being compensated.

C. Research Methodology

Because the relationship between law and computer science is complex, this paper utilizes a mixed doctrinal/disciplinary method of legal research. The doctrinal part heavily depends on the study of primary legal sources. This involves an examination of international copyright instruments, including the Berne Convention and the WIPO Copyright Treaty, national copyright statutes, including the Copyright Act, 1957 of India and the Copyright Act of 1976 of the United States, and relevant precedent concerning copyright enforcement in digital and AI-related contexts.[3]

The interdisciplinary aspect is the critical relationship of digital forensics and computer science literature with the legal analysis. Secondary sources like computer science white papers (especially those related to data extraction and model memorization), tech law treatises, peer-reviewed law journal articles and forensic science methodologies are synthesized. This will allow legal analysis of evidentiary burdens to be based on computational fact, not on technological abstraction. The research is analytical and prescriptive, as its goal is to not only pinpoint the existing legal shortcomings, but also to provide with concrete procedural and forensic answers.

D. Literature Review

In just five years, the discourse on AI and copyright has expanded rapidly in both the academic and professional worlds. Nevertheless, a careful survey of the literature shows that an important bias exists towards substantive law, and that there is a notable absence of literature on the procedural and forensic mechanics of evidence.

[3] WIPO Copyright Treaty, Dec. 20, 1996; Copyright Act, 1957; Copyright Act of 1976, 17 U.S.C. §§ 101–122.

There are two "polarized camps" of legal scholarship today that are divided on the issue of substantive rights. The first camp, which is more technologically utilitarian, strongly advocates for the benefit of AI developers. Such scholars, including Mark Lemley, are often heard saying that such use for AI training is a very transformative "fair use" (per US copyright doctrine) or a "text and data mining" exception (per the EU copyright doctrine; Japan also has such an exception). They say that AI doesn't "copy" expression, but rather analyze facts, ideas, and statistical correlations that aren't protectable.

The other side, typically supported by traditional IP scholars and fans of the creators' rights movement, including Jane Ginsburg, claims that data scraping is no different from "rampant digital piracy." They believe that the purpose of a generative AI is to create competing expressive works, so the first ingestion cannot be considered fair or transformative.

This substantive debate is significant, but it often assumes that copying, ingestion, and training use can be proved with sufficient clarity. The 2025 decisions in *Thomson Reuters*, *Bartz*, and *Kadrey* show that the evidentiary problem is now central to AI-copyright adjudication. These cases suggest that the "black box" is not always impenetrable: internal records, discovery, acquisition logs, licensing communications, and dataset histories may provide direct or quasi-direct evidence of sourcing and use. At the same time, they reinforce the need for reliable forensic methods because plaintiffs may still face serious difficulty proving the presence of specific works in training data, the manner of acquisition, and the market consequences of model training.

This has been the case since AI firms intentionally withhold the origin of their training data to preserve trade secrets, creating an unfair litigation landscape, legal experts agree. In contrast, computer science research by Nicholas Carlini et al., in *Extracting Training Data from Large Language Models*, demonstrates that language models may, in certain circumstances, memorize and reproduce portions of their training data. However, the idea of applying these computer science insights to civil procedure and evidence has not been fully addressed in the academic literature. This paper aims to fill that very same

gap, and will examine the practicalities of digital forensics and how these can meet the evidential requirements of an intellectual property tribunal.

1. The 2025 AI-Training Decisions: Discovery, Internal Records, and the Evidentiary Burden

The literature and case-law survey must also account for the significant 2025 decisions in *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.*, *Bartz v. Anthropic PBC*, and *Kadrey v. Meta Platforms, Inc.* These rulings are directly relevant to the evidentiary problem examined in this paper because they show that courts are beginning to resolve AI-training disputes not merely through abstract assumptions about "access" or "substantial similarity," but through documentary evidence, internal records, data-acquisition histories, and discovery concerning how training material was sourced.

In *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.*, the District of Delaware rejected Ross Intelligence's fair-use defense in relation to its use of Westlaw headnotes to train a competing AI-powered legal-research tool. The importance of the ruling for this article is evidentiary as much as doctrinal: the case did not turn merely on a speculative inference that Ross had "access" to protected material, but on the record concerning how the training data had been acquired and used. The decision therefore qualifies the manuscript's claim that plaintiffs are necessarily confined to external forensic inference. Where discovery produces internal documentation of acquisition, copying, or training use, the black-box problem may be reduced, although not eliminated.

In *Bartz v. Anthropic PBC*, the Northern District of California drew an important distinction between the use of books for LLM training and the creation or retention of a broader library of pirated materials. The court held that Anthropic's training use was transformative and could qualify as fair use, but it did not treat the downloading and retention of pirated books in a central library as automatically excused by that training purpose. This distinction directly supports the paper's thesis that evidentiary precision matters: plaintiffs must establish not only that their works were present somewhere in a defendant's data environment, but also how those works were acquired, copied, retained, filtered, and actually deployed in training datasets.

Kadrey v. Meta Platforms, Inc. further demonstrates that AI-training disputes may turn on the evidentiary record rather than on any categorical rule that training is always lawful or always infringing. The court granted Meta summary judgment on the fair-use defense on the record before it, but the reasoning emphasized the plaintiffs' failure to produce sufficient evidence of market harm. For a paper on evidentiary standards, this is a central development: even where unauthorized acquisition or use of copyrighted books is alleged, plaintiffs must still develop admissible evidence connecting the challenged copying to legally cognizable harm, including substitution, licensing-market injury, or other effects on the potential market for the works.

IV. RESEARCH & ANALYSIS

A. The Doctrinal Obsolete of Traditional Evidentiary Standards

The only way to understand the legal conundrum is to examine the familiar dynamics of copyright action. Courts universally rely on circumstantial evidence, as the type of evidence used to show copying can be circumstantial only if there is no direct evidence, such as a witness who claimed to have seen the defendant copy a file. To win a copyright case, the plaintiff has to prove the defendant had "access" to the copyrighted work and that the defendant's work shows "substantial similarity" to the plaintiffs.

With the digital internet, proving the "access" has become relatively easy. Courts have recognised that widespread online availability may support an inference of access where there is a reasonable possibility that the defendant encountered the copyrighted work. As AI developers are using automated web crawlers (like Common Crawl dataset) which are scraping billions of URLs, it's easy to prove, legally, that an AI firm had access to a publicly posted article, image, or codebase.

The second prong is the systemic failure: "substantial similarity." Under traditional jurisprudence, the "ordinary observer" test is used. A layman viewing both works would be led to believe that the defendant's work was an unauthorized use of the plaintiff's protectable expression? This test has been completely overcome by the algorithms of Generative AI. A Large Language Model that is trained on a copyrighted fantasy book

will not regurgitate the fantasy book word for word. Rather, it takes up the syntactic structures, word frequency and thematic patterns. Upon being asked, it can produce a mathematically new sequence of words, which is not a plagiarism of the original author's words, but completely depends on the value generated from the work of the original author.

AI output will not be a verbatim duplication of one item in the training set since AI generates and interpolates ideas, not copy/paste exact duplicates. Thus, using an output comparison to demonstrate input infringement is a fundamentally unsound legal strategy. The infringement was during the hidden training phase, when the production of a work into a server memory to obtain data from the work was made without permission; however, the traditional legal test only applies to the output, which was not the infringing one. The legal doctrine is thus poorly geared to the technological reality.

B. The 'Black Box' Problem and the Corporate Trade Secret Shield

Without the output being able to establish that it infringes, the plaintiff's only remedy is to look at the input the training data itself. In a civil action, however, determining what is contained within a training dataset is a procedural nightmare because of the opacity and intentional minimization of the same, known as the "black box" problem in AI.

AI models are extremely complex matrices of parameters sometimes with trillions of numbers. Even the developers who develop these models can't assert a relationship between a particular output and a particular input node. More seriously, AI developers are more likely to fight to prevent disclosure of their training datasets in the discovery stage of litigation. They often claim that training data they use are highly valuable, proprietary, commercial information, which gives them a competitive edge, and they invoke the protection of trade secret.

Developers intentionally create evidentiary bottlenecks by presenting the data as a proprietary trade secret. They require plaintiffs to request discovery at the time of filing, a motion for which prolific redactions, obfuscation, or assertions that the original datasets

have been lost, replaced, etc. by the abstracted algorithmic weights, are often met. These imbalances and prejudices the litigation process. Plaintiff is bound to prove unauthorized use, but is precluded from accessing the only means of discovery (the dataset or server logs) where the direct evidence of such use is located from doing so by any means legally allowed by corporate confidentiality doctrines.

C. Digital Forensics Methodologies as the Evidentiary Solution

For plaintiffs and IP courts to pierce through the black box and break through the trade secret barrier, there is a significant shift to more sophisticated digital forensics required. There are multiple methodologies that can be used to track digital footprints and determine the ingestion of a proprietary data set without the need for complete access to that data set.

1. **Hash Matching and Dataset Tracing:** Although there are some high-tech datasets that developers keep private, many basic AI models are based on open-sourced collections of data scraped from the web, like the LAION-5B network of images or the Books3 collection of texts. These are open sources that can be used by digital forensic experts with cryptographic hash functions (such as SHA-256). The experts can use this hash to conduct a comparative forensic scan of the known architecture of these huge scraping databases and to identify with certainty that the exact same digital file existed in the corpus that the AI developer used. This is a digital way of showing proof of ingestion.
2. **Data Extraction Attacks and Memorization:** The literature on machine learning has shown that LLM and diffusion models can exhibit the problem of "overfitting" that is, they can memorize specific snippets of their training data and not be able to generalize concepts. "Data extraction attacks" are the deliberate creation of very specific, malicious prompts that are intended to force the AI to spit out exactly what it was trained to regurgitate: text or watermarked images. For instance, if an AI generates the next sentence of a copyrighted, paywalled article exactly as it reads in that article, then this

constitutes a clear and convincing (definitive) forensic evidence that the article was fed into the model's training data.

D. The New York Times Company v. Microsoft Corporation, OpenAI, Inc. et al.: Use of Evidence at the Pleading Stage

A significant contemporary example of memorization-based pleading is the complaint filed in the Southern District of New York in *The New York Times Company v. Microsoft Corporation, OpenAI, Inc. et al.*, No. 1:23-cv-11195, before Judge Sidney H. Stein on 27 December 2023. The Times included worked examples alleging that ChatGPT outputs reproduced substantial portions of its articles, with overlapping text compared against the original works. The method resembled the data-extraction technique discussed above: prompts based on portions of protected articles were used to test whether the model would generate substantially similar or overlapping text. The pleading therefore did not rely solely on an inference of ingestion from stylistic similarity; it attempted to use model outputs as circumstantial evidence that protected works had been included in or learned from the training process.

The response to that evidence from OpenAI is instructive as well. The company described memorisation as a "rare defect" and said that the Times used its creation to thousands of times only to make people regurgitate, but failed to do so under the terms of use. The disagreement, in other words, was now about the reliability of the forensic protocol – what was the number of attempts made, who developed it, who it could be reproduced by, on what settings. These are the questions that a court may ask about the methodology of an expert under a reliability standard and at a pleading stage in an AI copyright case.

On 4 April 2025, Judge Stein substantially denied the defendants' motions to dismiss. The court allowed certain copyright infringement claims involving conduct outside the three-year limitation period to proceed, rejected dismissal of the contributory copyright-infringement claims, and allowed the trademark-dilution claims in the related *Daily News* action to continue. The passage should refer throughout to copyright infringement, not

patents or patent infringement. The court did, however, dismiss several claims under DMCA § 1202(b), while allowing other aspects of the copyright litigation to proceed. This distinction is important because memorization or output-overlap evidence may support some copyright theories, but it does not automatically satisfy every statutory element required for a DMCA copyright-management-information claim.

This division is important to the argument advanced here. A sufficient body of extraction-style or output-overlap evidence may assist a claim framed in terms of unauthorized reproduction or training-related copyright infringement, but such evidence must still be linked to the specific legal elements of each pleaded cause of action. Evidence of memorisation is not admissible as proof of memorisation to every cause of action, but only as evidence to prove the element to which it is offered. The discovery proceeding underway will test whether the forensic methods documented in this paper can withstand a contentious examination, and will be the first major judicial scrutiny of training data composition and acquisition practice involving the consolidated copyright actions against OpenAI and Microsoft.

E. Auditing APIs or Metadata Analysis

Evidence can also be found with the use of a forensics process that involves looking at the logs of the creator's own server. Forensic investigators can uncover the distinctive characteristics of popular AI data scrapper tools like GPTBot or CCBot by analyzing their server access records, network activity, and API usage. Analyzing server logs, network traffic, and API usage can reveal the specific digital signatures, IP addresses, and user-agent strings used by known tools such as GPTBot or CCBot. When a plaintiff can prove that a scraper, used to download protected directories, has been following a forensic trail, it is powerful circumstantial evidence of unauthorized reproduction. This is an irregular course focused on the evidentiary burden. This is an irregular course that will discuss the evidentiary burden.

The challenges with evidence are further complicated by transnationality of digital scraping. A U.S.-based AI firm can install servers in E.U. to access information stored in India.A U.S. based AI firm can put these servers in the European Union to collect

information stored in India. The question of jurisdiction in relation to the copying of documents (and hence the applicable copyright laws and evidentiary rules) is highly complicated.

The EU Digital Single Market Directive creates specific exceptions for text and data mining, including a broader exception where the rightsholder has not expressly reserved rights in an appropriate manner.[4]In this setting, the burden of the forensics changes. The plaintiff needs to not only show that the AI scraped his work, but that his digital file actually had the correct "opt-out" metadata at the exact time it was scraped. This demands that metadata states be preserved over time through means of tools such as the Wayback machine, or cryptographic time-stamping, and introduces another layer of procedure into litigation.

V. SUGGESTIONS AND RECOMMENDATIONS

Today, copyright law and AI are incompatible. The onus is on the disenfranchised creator, and the evidence is completely controlled by the AI developer. This paper proposes the following changes to the law and procedures to restore equity and make copyright survive in the digital age:

1. **Required Data Disclosures, EU AI Act Transparency, and Training Logs:** Legislatures around the world need to reform the procedural rules governing intellectual-property disputes involving machine learning. This recommendation is no longer merely theoretical. Article 53(1)(d) of the EU AI Act already requires providers of general-purpose AI models to draw up and make publicly available a sufficiently detailed summary of the content used for training, in accordance with a template provided by the AI Office. This mechanism is an important regulatory precedent because it recognises that

[4] 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, arts. 3–4.

copyright enforcement in the AI context requires some level of training-data transparency. However, a public summary of training content may still be insufficient for litigation purposes where the issue is whether a particular work was ingested, copied, retained, filtered, or used in training. Accordingly, AI developers should also be legally required to maintain secure and auditable training logs recording metadata, source URLs, dataset identifiers, acquisition dates, filtering steps, and hash values of training materials. Such logs need not be made fully public, but should be discoverable under judicial supervision where a plaintiff establishes a plausible infringement claim, subject to protective orders preserving trade secrets and confidential commercial information.

2. **Implementation of Forensic "Watermark" Standards:** Legislative and judicial institutions should formally acknowledge and establish cryptographic watermarking and tracking of forensic metadata. Once a creator puts a standardized, machine-readable copyright claim into their digital product, the removal and/or stripping of that metadata by an AI scraping company should be considered copyright management information violation and should result in full liability, independent of the substantial similarity test.
3. **Presumption of Ingestion:** Courts should apply the rebuttable "presumption of ingestion" to civil litigation involving large-scale AI models with training sets that include large amounts of the internet. The plaintiff should be entitled to shift the burden of proof to the AI developer to provide forensic proof, such as on their server logs, that the specific plaintiff's work was removed or filtered from the dataset, if it can be shown that it was widely available online during the period of time the plaintiff's copyrighted work was being trained on.
4. **Creation of Specialized IP-Tech Tribunals:** These forms of data extraction attack, hash matching, and neural network architectures are highly technical and if used as evidence in a standard civil court they are likely to be more difficult to evaluate. Sole or specialized intellectual property tribunals comprised of both legal and forensic computer scientists should be set up by

jurisdictions to precisely hear and resolve disputes on copyright ownership of AI.

5. **EU AI Act as a Regulatory Precedent:** The EU AI Act should be expressly acknowledged as the leading real-world example of training-data transparency regulation. Article 53(1)(c) requires providers of general-purpose AI models to put in place a policy to comply with EU copyright law, including respect for rights reservations under Article 4(3) of Directive (EU) 2019/790, while Article 53(1)(d) requires publication of a sufficiently detailed summary of training content. Together, these provisions link copyright compliance, text-and-data-mining opt-outs, and AI-training transparency. Nevertheless, the EU model should be treated as a baseline rather than a complete evidentiary solution. A summary template may assist authors, regulators, and courts in understanding the broad categories of training content, but it may not provide the work-level proof required to establish infringement in adversarial litigation.
6. **Limits of Summary-Based Transparency:** The EU training-content summary template strengthens the policy case for transparency, but it does not eliminate the forensic "black box." For infringement litigation, the relevant evidentiary question is often not whether a model was trained on broad categories such as books, news, code, images, or web text, but whether the plaintiff's specific protected work was copied or retained in the relevant dataset at the relevant time. Therefore, this paper's proposed training-log requirement remains necessary even under the EU regime. Public summaries should be supplemented by confidential, court-supervised access to auditable records capable of verifying dataset provenance, rights-reservation compliance, deduplication, filtering, and removal.

VI. CONCLUSION

It is not only that AI tools threaten copyright law by enabling a process of copying and appropriating intellectual work without the creator's consent, it is also that the very structures of the legal system threaten to undermine the very foundation of copyright

which is that the creator should be able to control and benefit from his or her intellectual work. Enforcement of the copyright law is as firm as the evidence that can be presented to establish infringement. As long as the "substantial similarity" test remains in place, it can be manipulated by AI developers to circumvent their IP obligations, while relying on the "black box" of trade secrets.

Plaintiffs can begin to crack this algorithmic opacity by aggressively weaving in more rigorous digital forensic methods including data extraction attacks, data dataset hashing, and metadata auditing into the legal process. But it is no replacement for strong statutory change. Legislators cannot afford to be passive when it comes to the matter of AI authorship: they should be proactive and address the reality of the situation. Incorporating the forensic standards into civil procedure, presumptions of ingestion, and dataset transparency are key steps. This interdisciplinary approach is essential to enable the compensation of human authors for their efforts and to ensure the integrity, enforceability and constitutional purpose of copyright law in the current highly automated and opaque digital world.

VII. REFERENCES

1. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). *Extracting Training Data from Large Language Models*. In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.
2. The Copyright Act, 1957 (Act No. 14 of 1957), §§ 2(d), 14, India.
3. The Copyright Act of 1976, 17 U.S.C. §§ 102, 107, United States.
4. 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.
5. Ginsburg, J. C. (2003). The theory of authorship for comparative copyright law. *DePaul Law Review*, 52, 1063.
6. Lemley, M. A., & Casey, B. (2020). Fair learning. *Texas Law Review*, 99, 743.

7. Complaint, *The New York Times Company v. Microsoft Corporation, OpenAI, Inc. et al.*, Case No. 1:23-cv-11195 (S.D.N.Y. 2023).
8. *Andersen v. Stability AI Ltd.*, No. 23-cv-00201-WHO (N.D. Cal. 2023).
9. *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).
10. World Intellectual Property Organization (WIPO). (1886). *Berne Convention for the Protection of Literary and Artistic Works*.