ISSN: 2837-4509

Iris Online Journal of Arts and Social Sciences

DOI: 10.33552/IOJASS.2025.02.000542

Iris Publishers

Review Article

Protecting the Candy: A Case for Data and Software Rights in the Age of Machine Learning

Charles R. Macedo*[1]

Partner, Amster, Rothstein & Ebenstein, LLP, Chloe Vizzone, Associate, Amster, Rothstein & Ebenstein, LLP, and ChatGPT 4o

Corresponding author: Charles R. Macedo, Partner, Amster, Rothstein & Ebenstein, LLP, Chloe Vizzone, Associate, Amster, Rothstein & Ebenstein, LLP, and ChatGPT 4o

Received Date: May 28, 2025

Published Date: June 14, 2025

Abstract

As machine learning and generative AI technologies reshape innovation and creativity, legal frameworks for intellectual property (IP) have lagged behind. This article argues that data and software, the fundamental ingredients of AI systems, deserve dedicated legal protections analogous to physical goods. Using the analogy of a candy store, this piece frames generative AI outputs, platforms, and data preparation as products and supply chains deserving IP rights. A new category of "**Data Rights**" Ⓓ and "**Software Rights**," Ⓢ grounded in originality and commercial investment, should be introduced to fill the current gap. These rights would be simple to enforce, valid for 15 years, and provide statutory royalties to all contributors, from data curators to platform developers.

[1] Charles R. Macedo is a Partner at Amster, Rothstein & Ebenstein LLP where his practice focuses on all facets of intellectual property with a special emphasis on computer-implemented innovation. Chloe Vizzone is an associate at Amster, Rothstein & Ebenstein LLP. ChatGPT 4o was used to generate the first draft of this article with prompting by Mr. Macedo, and further editing by Mr. Macedo and Ms. Vizzone. ChatGPT 4o is not an author under current legal standards.

Introduction

Imagine walking into a candy store, admiring the shelves lined with bright wrappers and mouthwatering sweets, and then simply walking out with your favorite chocolate bar-without paying. Most would agree such an act is theft. Under state criminal codes, this would typically constitute petty larceny or shoplifting, with penalties including fines or jail time. Yet in today's AI-driven economy, analogous behavior is not just common; it is often legal. Developers, platforms, and users routinely exploit data and generative AI outputs without attribution or compensation to those who created, curated, or maintain the underlying resources. As the economic and creative value of machine learning tools continues to surge, so too must the legal protections afforded to their building blocks.

Problem: Outdated IP Laws in a Post-ChatGPT Economy

The rise of ChatGPT in late 2022 marked a paradigm shift in public awareness and commercial adoption of artificial intelligence ("AI") and generative technologies. Within months, generative AI tools were being used in virtually every sector-from education and entertainment to scientific research and legal services. This explosion of usage has propelled generative AI to become a critical engine of the digital economy, projected to contribute trillions of dollars to global GDP in the coming decade.

Yet existing legal frameworks-primarily copyright and patent law-are ill-suited (at least in the U.S.) to protect the core elements of these systems: data and software.

Copyright law struggles to protect software-generated content due to the "authorship" requirement, which excludes outputs not created by a human author. Additionally, raw and curated datasets used in training AI generally do not qualify for copyright protection unless they exhibit a creative selection or arrangement leaving most training data unprotected.

While a stolen candy bar invokes clear criminal consequences, an AI-generated image copied without attribution may go unpunished.

Likewise, while software code can be copyrighted, functionality and architecture often fall outside its scope, and patent law imposes a high bar for novelty and non-obviousness, and has been reluctant to find software patent-eligible.

This legal vacuum creates significant risks:

i. Developers lack enforceable rights over generative outputs, undermining monetization and investment.

ii. Platforms cannot ensure exclusive use of their software systems or training pipelines.

iii. Data contributors and curators go uncompensated, weakening incentives for high-quality data preparation.

In sum, the current IP regime, built for a pre-AI world, fails to recognize or reward the layered economic and creative contributions required to build machine learning systems.

More critically, these legal constructs are rooted in the economic and technological paradigms of the 18th and 19th centuries. The foundational structure of copyright was designed to incentivize and protect authors of literary and artistic works in a print-dominated society. Similarly, patent law emerged to promote invention during the Industrial Revolution, where mechanical processes and chemical compounds formed the backbone of economic value. These laws were well-suited to the tangible, human-centered outputs of their time.

But the 21st century operates on an entirely different substrate: information. The raw materials of today's economy are no longer cotton, coal, or steel-but data, algorithms, and computation. Intellectual labor is increasingly performed not only by humans but also by intelligent machines. Creation is less about fixing ink to paper and more about training models on vast datasets, producing outputs through layers of computation, and refining them through human-machine collaboration. Despite this, our IP regime continues to hinge on antiquated notions of authorship, and fixation.

As AI systems generate images, text, and software that rival or exceed human creations, legal questions of ownership, attribution, and compensation have become more urgent. Traditional IP laws do not answer these questions adequately because they were never designed to accommodate non-human creators, or the distributed, iterative nature of machine learning development. Without reform, we risk suppressing innovation, misallocating value, and fostering inequity across the AI economy.

In this context, reform is not only with is inevitable. Just as the law adapted to the printing press, the camera, and the internet, it must now evolve to embrace the realities of generative AI. Doing so will require new legal instruments that account for the complexity, scale, and economic significance of digital creation. It will also demand a reimagining of what it means to protect an "original work" or an "inventive step" in a world where machines contribute to both.

We propose that this reform should begin with the establishment of two new IP frameworks: **Data Rights** and **Software Rights**. These rights would recognize the economic and creative value of curated data and model design. They would offer protections where existing regimes fall short and provide a balanced, time-limited structure for attribution, licensing, and fair remuneration. They represent a first step toward reconciling law with the logic of the digital age.

The Output: The Candy on the Shelf

The final product of a generative AI system-whether it is a poem, image, software code, or synthetic data-is akin to a candy bar on a shelf. Just as a consumer cannot lawfully walk into a candy shop and take a bar of chocolate without paying, users should not be able to freely extract and commercialize AI-generated content.

This principle is rigorously upheld in the physical world. Criminal statutes, such as New York Penal Law § 155.25, treat

the unauthorized taking of tangible goods-including candy-as "petit larceny". Retail theft is monitored by surveillance systems, prosecuted by district attorneys, and deterred through fines, community service, and incarceration. In contrast, the digital appropriation of AI-generated works often goes unchecked due to gaps in intellectual property law.

Under the Copyright Act of 1976, copyright protection is limited to "original works of authorship fixed in any tangible medium of expression". The U.S. Copyright Office has clarified that works created without human authorship are ineligible for copyright, as articulated in its 2023 guidance on AI-generated works. This limitation leaves AI-generated content unprotected, even when it has substantial economic or creative value.

Legal scholars such as Prof. Jane Ginsburg have emphasized that the concept of authorship is central to copyright, and without a human agent, courts are reluctant to extend protection. As a result, the moment an AI-generated poem, image, or article is released, it can be copied and redistributed without fear of legal repercussions. This asymmetry distorts market incentives and disincentivizes creators from investing in AI development.

The Platform: The Storefront That Sells the Candy

The generative AI platform is not unlike the neighborhood candy store: a curated, maintained, and monetized environment designed to offer products to the public. These platforms represent significant capital outlay-often in the tens or hundreds of millions of dollars-for computing resources, engineering teams, interface development, cybersecurity infrastructure, marketing strategies, and compliance.

In the brick-and-mortar world, protections for business operators are well established. Trademark law under the Lanham Act protects the visual identity, slogans, and branding of a store from infringement and dilution. Commercial landlords and retail franchises benefit from well-defined leasing contracts, trade dress protections, and franchise laws.

In the digital realm, however, these safeguards are patchy. While platform names and logos may be trademarked, the user interface design, the recommendation engines, and the model architectures behind generative platforms enjoy minimal protection.

Copyright law explicitly excludes protection for "ideas, procedures, processes, systems, methods of operation, concepts, principles, or discoveries," which limits the protection of AI system functionalities.

Trade secret law, such as that codified in the Defend Trade Secrets Act of 2016, offers some remedies, but these are difficult to enforce internationally and provide no recourse once a secret is made public or reverse-engineered. As many legal scholars and AI experts have pointed out, trade secrets rely on confidentiality, which clashes with the transparency ethos of open science and responsible AI development.

The Data: The Ingredients That Make the Candy

Just as no candy bar can be made without ingredients, no AI model can exist without data. Training data serves as the sugar, cocoa, and milk of the machine learning pipeline. It is the essential fuel from which patterns are learned, and without it, there is no model.

Candy manufacturers pay for their ingredients and are legally bound by contracts and supply chains governed by the Uniform Commercial Code (UCC). A breach of contract results in clear civil liability; theft or misappropriation can result in criminal charges or trade secret litigation.

In contrast, training datasets are often compiled through web scraping, API extraction, or bulk licensing from undisclosed sources. While some uses may fall under fair use exemptions (as discussed in *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015)), this doctrine is narrow and context-dependent. Moreover, many creators object to the use of their content for AI training, especially when it competes with or displaces their own economic activity.

Legal commentators such as Mark Lemley have argued that existing copyright law lacks the doctrinal tools to address non-consumptive, large-scale data use. The sui generis database protection offered in the EU under the Database Directive (96/9/EC) has no parallel in the United States, creating a regulatory arbitrage problem. Without an American counterpart, curated datasets-especially those requiring human labor to annotate, structure, and clean-remain largely unprotected.

The Data Preparation: The Candy-Making Process

Turning raw sugar and milk into a finished candy bar requires more than a recipe-it requires skilled labor, machinery, sanitation, packaging, branding, and logistics. These processes are protected by a variety of legal and contractual rights in the physical economy.

Similarly, data must be curated cleaned, labeled, normalized, enriched, and formatted before it becomes usable in training a model. These processes are labor-intensive, often requiring data scientists, annotators, domain experts, and engineers. For example, the ImageNet project involved hundreds of thousands of human hours to label images accurately-work underpins countless AI applications today.

However, traditional IP frameworks fail to recognize these contributions. As the Supreme Court held in *Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340 (1991), mere investment in data collection and compilation does not meet the originality requirement of copyright. While this ruling reinforced the idea-expression dichotomy, it left producers of non-expressive data vulnerable.

Scholars such as Jerome Reichman and Paul Uhlir have proposed sui generis protection for data compilation based on investment and utility rather than creativity. Such proposals have gained little

traction in the U.S., but the need is growing more acute as AI models increasingly depend on curated, domain-specific datasets.

Proposing New Rights: The Case for Data Rights and Software Rights

To resolve these challenges, the U.S. should introduce two new IP categories: **Data Rights** and **Software Rights**.

Data Rights would protect original, curated datasets-especially those demonstrating selection, coordination, or investment. These rights would:

i. Last for 15 years from first commercial use.

ii. Prohibit unauthorized copying, re-use, or distribution.

iii. Allow for reasonable licensing and fair royalties.

iv. Be subject to clear, simple criteria for enforcement (e.g., registration and publication).

This concept mirrors the EU Database Directive, but would be tailored to the American legal landscape. Instead of requiring creativity, it would recognize investment, labor, and organization. This would ensure that data compilers-especially in sectors like healthcare, scientific research, and education-receive returns on their efforts.

Software Rights would protect machine learning systems and underlying algorithms, similar to design patents, including:

i. The structure, training architecture, and tuning of AI models.

ii. A 15-year term, renewable once for systems under active use.

iii. Enforcement mechanisms modeled after copyright and trade secret law.

Software Rights would be narrower than full patent protection but broader than copyright in covering functional elements of AI systems. They could adopt principles from the Semiconductor Chip Protection Act of 1984, which recognizes the industrial design of microchips without demanding full patent standards. By protecting the model architecture as an engineered system, these rights would balance innovation incentives with competition and interoperability.

These rights would ensure that developers and data curators are fairly compensated while still allowing access under fair terms. They would provide legal certainty, promote responsible investment, and reward those who contribute to the AI ecosystem-without stifling downstream innovation.

Acknowledgement

None

Conflict of Interest

No conflict of interest.

1. Introducing ChatGPT, Open AI (Nov. 30, 2022), https://openai.com/index/chatgpt/.
2. Global Future Counsel on International Trade and Investment, White Paper, ChatWTO: An Analysis of Generative Artificial Intelligence and International Trade, World Economic Forum (Sept. 4, 2024), https://www.weforum.org/publications/chatwto-an-analysis-of-generative-artificial-intelligence-and-international-trade/#:~:text=Generative%20artificial%20intelligence%20could%20contribute,reshaping%20industries%20and%20international%20trade ("Generative artificial intelligence could contribute an estimated $4.4 trillion annually to the global economy, reshaping industries and international trade.").
3. In Thaler v. Perlmutter, 130 F.4th 1039, 1041 (D.C. Cir. Mar. 18, 2025), the D.C. Circuit concluded "[t]he Creativity Machine cannot be the recognized author of a copyrighted work because the Copyright Act of 1976 requires all eligible work to be authored in the first instance by a human being."
4. See Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340, 344 (1991) ("[F]acts are not copyrightable; [but] compilations of facts generally are.")
5. See 17 U.S.C. § 102(b) ("In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.") (emphasis added); Baker v. Selden, 101 U.S. 99 (1879) (copyright law protects expression not underlying idea in a work).
6. 35 U.S.C. § 102.
7. 35 U.S.C. § 103.
8. 35 U.S.C. § 101. See, e.g., Recentive Analytics, Inc. v. Fox Corp., 134 F.4th 1205, 1211 (Fed. Cir. 2025) ("This case presents a question of first impression: whether claims that do no more than apply established methods of machine learning to a new data environment are patent eligible. We hold that they are not.").
9. See generally, A Brief History of Copyright in the United States, Copyright.gov, https://www.copyright.gov/timeline/#:~:text=To%20promote%20the%20Progress%20of,their%20respective%20Writings%20and%20Discoveries.%E2%80%9D.
10. See generally, A Bill to Promote the Progress of the Useful Arts (the Patent Act), H.R. 41, 1st Cong. (1790).
11. 17 U.S.C. § 102(a).

12. Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 37 C.F.R. Part 202 (2023).

13. Jane C. Ginsburg, The Concept of Authorship in Comparative Copyright Law, 52 DEPAUL L. REV. 1063 (2003).

14. 15 U.S.C. § 1125.

15. 17 U.S.C. § 102(b).

16. 18 U.S.C. § 1836, et seq.

17. See, e.g., Adnan Masood, PhD., Intellectual Property Rights and AI-Generated Content – Issues in Human Authorship, Fair Use Doctrine, and Output Liability, Medium (Apr. 4, 2025), https://medium.com/@adnanmasood/intellectual-property-rights-and-ai-generated-content-issues-in-human-authorship-fair-use-8c7ec9d6fdc3.

18. Best of 2023: Copyright Law & Artificial Intelligence: Is Training AI With Other's Data Fair Use – Professor Mark Lemley (Stanford Law), TECHNICALLY LEGAL (Dec. 28, 2023), https://www.tlpodcast.com/professor-mark-lemley-of-stanford-explains-why-he-thinks-that-copyrighted-works-used-to-train-ai-fall-should-under-the-fair-use-exception-to-copyright-law/.

19. Database Directive 96/9, 1996 O.J. (L 77).

20. https://www.image-net.org (as of March 11, 2021 ImageNet has 14,197,122 images, 21841 synsets indexed).

21. Jerome H. Reichman & Paul F. Uhlir, Database Protection at the Crossroads: Recent Development and Their Impact on Science and Technology, 14 BERKELEY TECH. L. J. 793 (1999).

22. 17 U.S.C. §§ 901–914.