TRANSFORMATIVE USE, THE BLACK BOX PROBLEM, AND THE UNFINISHED ARCHITECTURE OF AI COPYRIGHT LAW

Akshit Mathur, Jindal Global Law School

ABSTRACT

The doctrine of transformative fair use has served as copyright law's principal mechanism for accommodating technological change. From the photocopier to the internet, courts have repeatedly held that repurposing copyrighted material for a fundamentally different function can excuse what would otherwise constitute infringement. Authors Guild, Inc. v. Google Inc. (2d Cir. 2015) represented the high-water mark of this approach: mass digitization of millions of books was deemed transformative because courts could inspect, measure, and verify the scope of Google's use through its snippet-display interface. A decade later, Bartz v. Anthropic PBC (N.D. Cal. 2025) asked whether the same logic could extend to large language models trained on millions of copyrighted works — and the answer revealed a doctrinal fault line that Authors Guild had papered over. This article argues that the two cases, read together, expose a structural vulnerability in the fair use framework: the doctrine can only function as intended when courts can empirically verify that a claimed transformation is genuine rather than merely asserted. In the age of opaque generative AI, that verification has become impossible. The article proceeds in three parts. Part I analyses the Authors Guild precedent and argues that its reasoning rested, implicitly, on the inspectability of Google's use. Part II examines how Bartz both extended and strained that precedent, with particular attention to the 'black box problem' — the inability of courts, authors, or even developers to trace how copyrighted material shapes model behaviour. Part III proposes a governance framework centred on mandatory dataset disclosure, third-party auditing, and collective licensing, arguing that these mechanisms are not merely desirable policy supplements but necessary prerequisites for the fair use doctrine to operate coherently in the context of generative AI.

I. INTRODUCTION: A DECADE, A DOCTRINE, AND TWO CASES

In October 2015, the Second Circuit Court of Appeals decided *Authors Guild, Inc. v. Google Inc.*, holding that Google's mass digitization of over twenty million copyrighted books constituted transformative fair use under Section 107 of the U.S. Copyright Act.[1] The decision was widely celebrated as a landmark for digital scholarship and a vindication of the principle that copyright law must adapt to serve its foundational purpose, the promotion of knowledge, even when that adaptation requires tolerating large-scale, unauthorized reproduction of protected works.

The legal test applied by the Second Circuit was the four-factor framework crystallized in *Campbell v. Acuff-Rose Music, Inc.*[2] the purpose and character of the use, including whether it is transformative; the nature of the copyrighted work; the amount and substantiality of the portion used; and the effect of the use on the potential market for the original.[3] What is less frequently noted, however, is that the Second Circuit's confidence in applying each of these factors rested on a crucial empirical condition: the court could see what Google had done. It could inspect the snippet-view interface, verify that no more than three non-consecutive lines were displayed, confirm that technological safeguards prevented reconstruction of full texts, and evaluate the market effect through evidence of enhanced book discoverability. Transformativeness, in *Authors Guild*, was not merely asserted, it was demonstrated.

Fast-forward ten years. In June 2025, Judge William Alsup of the Northern District of California decided *Bartz v. Anthropic PBC*, a copyright suit brought by authors Charles Graeber, Kirk Wallace Johnson, and Andrea Bartz against Anthropic, the developer of the large language model Claude.[4] The plaintiffs alleged that Anthropic had ingested over seven million digitized books, including, crucially, pirated copies obtained through repositories such as Books3, LibGen, and PiLiMi, to train its models. Judge Alsup held that training on legally acquired books was "quintessentially transformative," since the works were converted into non-expressive statistical representations rather than reproduced for their expressive content. He refused, however, to extend fair use protection to pirated materials. Anthropic subsequently settled for $1.5 billion, compensating affected authors at approximately $3,000 per work.

[1] Authors Guild, Inc. v. Google Inc., 804 F.3d 202 (2d Cir. 2015).
[2] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).
[3] U.S. Copyright Act of 1976, 17 U.S.C. § 107.
[4] Bartz v. Anthropic PBC, No. 3:24-cv-05417-WHA (N.D. Cal. June 23, 2025).

The ruling is significant not only for what it decided, but for what it could not verify. Unlike the *Authors Guild* court, Judge Alsup had no window into the model's operations. He could not inspect the internal representations that Anthropic's engineers had extracted from the plaintiffs' books. He could not confirm that those representations were truly "non-expressive," or that the model had not, in some meaningful sense, memorized and reproduced the creative architecture of the works it consumed. The court accepted transformativeness on the basis of a theoretical account of how large language models work, not on an empirical verification of what this particular model actually did with these particular books.

This article argues that the gap between those two epistemic positions, namely the inspectable use in *Authors Guild* and the opaque use in *Bartz*, is not merely a technical inconvenience. It is a structural vulnerability in the fair use doctrine itself. Part II analyses the *Authors Guild* precedent in detail, arguing that the court's reasoning was implicitly conditioned on the verifiability of Google's use. Part III examines how *Bartz* extended that precedent into territory where its foundational assumptions no longer hold, and identifies the specific ways in which the black box problem distorts each of the four fair use factors. Part IV proposes a governance framework comprising mandatory dataset disclosure, third-party auditing, and collective licensing as the necessary infrastructure for the doctrine to operate coherently in the age of generative AI.

II. AUTHORS GUILD V. GOOGLE: TRANSFORMATIVENESS AND THE INSPECTABLE USE

A. The Facts and the Four-Factor Analysis

Google's Library Project, launched in partnership with major research libraries, involved the wholesale digitization of millions of copyrighted books. The resulting Google Books database allowed users to search for terms and receive "snippet views," excerpts of approximately three lines, from books containing those terms.[5] The Authors Guild challenged this as copyright infringement, arguing that the digitization itself constituted an unauthorized reproduction and that the snippet view created an unauthorized derivative work.

The Second Circuit disagreed on all counts. On the first factor, purpose and character, the court

[5] Vanessa Campbell, Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015), 27 DePaul J. Art, Tech. & Intell. Prop. L. 1 (2016).

held that Google's use was highly transformative: it had converted the expressive content of books into a research and discovery tool, adding new utility without supplanting the original communicative function of the works.[6] Relying on *Campbell*, the court confirmed that commercial motivation did not preclude a finding of transformativeness, provided the new purpose was sufficiently distinct from the original.

On the second factor, the nature of the copyrighted work, the court acknowledged that the books were creative works entitled to strong protection but gave this factor less weight given the strength of the transformativeness finding. On the third factor, the court found that Google's display of snippets, limited in length and fragmented by design, did not constitute a substantial use of the works. Technological safeguards specifically prevented users from manipulating search queries to reconstruct full texts.[7]

On the fourth and most economically significant factor, the effect on the market, the court held that Google's use was unlikely to substitute for the original works. The snippet view was deliberately incomplete; it directed users toward, rather than away from, the books themselves, performing a discovery function that the court regarded as beneficial to authors' markets rather than harmful.

B. What the Ruling Implicitly Required: Verifiability

The *Authors Guild* decision is conventionally read as a case about digitization, public access, and the scope of transformative use. But there is a less-remarked precondition embedded throughout the court's analysis: every factual determination the court made rested on evidence it could examine directly. The snippet-view interface was demonstrably limited. The technological safeguards were documented and verifiable. The market effects could be studied through sales data and library usage statistics.

This verifiability was not incidental. The four-factor test as articulated in *Campbell* is fundamentally an empirical inquiry: courts are asked to assess what was actually done with the copyrighted material, how much of it was used, and what effect that use actually had on the market.[8] In *Authors Guild*, the court could answer all of these questions with reasonable

[6] Authors Guild, 804 F.3d at 207.
[7] Authors Guild, 804 F.3d at 214.
[8] Campbell, 510 U.S. at 577–78.

confidence because Google's use was, in relevant part, observable. The transformative claim was not merely theorized but was evidenced by the product itself.

This point matters enormously for what follows. When courts in subsequent AI cases invoke *Authors Guild* as precedent for the proposition that training on copyrighted works is transformative, they are extracting a legal conclusion from a factual context that no longer obtains. The precedent was built on an inspectable use; it is being applied to an opaque one. That is not an extension of the doctrine; it is a category error.

III. BARTZ V. ANTHROPIC: THE BLACK BOX AND THE FOUR FACTORS

A. The Facts and the Ruling

Anthropic's alleged conduct in *Bartz* differed from Google's in two significant respects. First, a substantial portion of the training corpus was not lawfully acquired: Anthropic allegedly obtained pirated copies of books through online repositories, and destructively scanned purchased print copies to build an internal digital library. Second, and more fundamentally, the use of the copyrighted material was not display but ingestion. The books were fed into a machine learning pipeline and processed to extract statistical patterns. Once training was complete, the original texts were not shown to users, quoted in outputs, or made searchable in any conventional sense. They had, in theory, been converted into numerical weights distributed across billions of parameters.

Judge Alsup's ruling drew a sharp distinction between these two aspects of Anthropic's conduct. Training on legally acquired books, he held, was transformative because the works had been converted. from expressive content into non-expressive statistical representations, enabling a new functionality (human-like text generation) that bore no substitutive relationship to the originals.[9] The use of pirated materials, by contrast, could not be justified under fair use regardless of the underlying training process, because the unlawful acquisition and retention of the materials displaced legitimate licensing markets and could not be excused by the transformative nature of the downstream use.[10]

[9] Bartz, No. 3:24-cv-05417-WHA, slip op. at 14.
[10] Id. at 17.

B. The Black Box Problem: A Distortion Across All Four Factors

While the ruling's distinction between lawful and unlawful acquisition is coherent and defensible, the deeper problem identified in the scholarly literature, and which the settlement leaves unresolved, is that the court could not actually verify the transformativeness claim it was accepting.[11] This is the black box problem, and it distorts the fair use analysis at every stage of the four-factor test.

On the first factor, the court accepted Anthropic's account that training converts expressive works into non-expressive statistical representations. This is theoretically plausible because it describes, at a high level of abstraction, what machine learning pipelines are designed to do. But it is not an empirical finding about what *this model* actually did with *these books*. Large language models are not uniform: they differ in architecture, training procedure, and the degree to which they memorize versus abstract from their training data. Studies suggest that models exhibit varying degrees of verbatim memorization, particularly for texts that appear frequently in their training corpora.[12] Neither the court nor the plaintiffs had the tools to determine where on the memorizationabstraction spectrum Anthropic's models fell with respect to the specific works at issue.

On the third factor, amount and substantiality, the opacity problem is most acute. In *Authors Guild*, "amount" could be quantified: three lines per snippet, no consecutive passages, no full-text reconstruction. In *Bartz*, no equivalent quantification was possible. The relevant question, namely how much of each book is encoded in the model's parameters and in what form, has no tractable answer.[13] The model's internal representations are distributed, high-dimensional, and noninterpretable by any currently available method. The court could not determine whether Anthropic had used the "heart" of each work, meaning the qualitative core of the author's creative contribution, because no one can currently look inside the model and answer that question.

On the fourth factor, market effect, the opacity problem compounds. The *Authors Guild* court could assess market harm through the snippet interface: users who received three non-

[11] Yangzi Li, Does Black-Box Machine Learning Shift the U.S. Fair Use Doctrine?, 16 J. Intell. Prop. L. & Prac. 1175 (2021).
[12] Singh, supra note 6, at 3.
[13] Hassija et al., supra note 7, at 46.

consecutive lines of text were unlikely to treat that as a substitute for reading the book.[14] The market dynamics of a large language model are far more complex and unpredictable. A model trained on an author's books might generate outputs that closely approximate the author's distinctive voice, prose style, or narrative structures, effects that could suppress demand for the original works without ever quoting them verbatim.[15] These substitution effects are difficult to detect, difficult to attribute, and entirely invisible to a court conducting a fair use analysis without access to the model's internals.

The result is a fair use determination made under conditions of systematic uncertainty. The court accepted transformativeness because the theory of how LLMs work is plausible and because accepting it is, on balance, preferable to holding all AI training categorically infringing.[16] But that is a policy judgment dressed in doctrinal language. It is not the empirical fair use analysis that *Campbell* and *Authors Guild* contemplated, and it is not adequate governance for a technology with the cultural and economic reach of generative AI.

C. The Piracy Holding: A Right Answer for the Wrong Reasons?

The court's refusal to extend fair use to pirated materials is the most straightforward aspect of the ruling, and it is also the aspect most likely to have practical significance in future litigation. The logic is compelling: fair use is a defence against infringement, not a licence to acquire infringing copies.[17] An AI company that builds its training corpus by downloading pirated e-books cannot subsequently claim that the transformative nature of its training process retroactively launders the unlawfulness of its acquisition.

This holding is correct, but it does not resolve the deeper problem. The black box problem applies equally to models trained on lawfully acquired data, and the $1.5 billion settlement was premised on the piracy issue, not on a finding that the training process itself was impermissible. The settlement, and the ruling it follows, leaves open the question of how courts should adjudicate fair use claims where the training corpus was lawfully acquired but the model's internal operations remain opaque. That is the case that will define the next decade of AI

[14] Authors Guild, 804 F.3d at 216 (noting that the snippet view created a 'discovery mechanism' that benefited the book market rather than displacing it).
[15] Li, supra note 5, at 1182.
[16] Bartz, No. 3:24-cv-05417-WHA, slip op. at 19.
[17] Id. at 22.

copyright litigation.

IV. TOWARD A GOVERNANCE FRAMEWORK: MAKING THE BLACK BOX LEGIBLE

A. The Limits of Doctrinal Adjustment

The instinctive legal response to the black box problem is to adjust the fair use doctrine by shifting burdens of proof, developing new evidentiary presumptions, or creating safe harbours for AI developers who satisfy certain procedural requirements. These adjustments have value, but they do not address the root problem: the fair use doctrine cannot function as an empirical inquiry if the empirical facts are inaccessible.[18]

Courts cannot adjust their way out of this difficulty. The four-factor test requires factual findings, and factual findings require evidence. If the relevant evidence, namely how the model processed the copyrighted material, what representations it formed, and how those representations influence its outputs, is systematically unavailable, then courts are not applying the fair use doctrine: they are guessing at its conclusions. That is not a doctrinal problem. It is a governance problem, and it requires a governance solution.

B. Mandatory Dataset Disclosure

The first element of a workable framework is mandatory disclosure of training datasets. AI developers should be required to maintain auditable records of the materials used to train their models, including records that identify the works included, their sources, and the means by which they were acquired. This is not a novel proposal: analogous requirements exist in other regulated industries where the provenance of inputs has legal significance.[19]

Dataset disclosure serves two functions in the copyright context. First, it makes the first and third fair use factors tractable. Courts and rights holders can determine what was used, in what quantity, and whether the acquisition was lawful, the threshold questions that *Bartz* identified as determinative. Second, it creates the evidentiary record on which a genuine market-effect analysis can be based. Without knowledge of what is in the training corpus, rights holders

[18] Li, supra note 5, at 1184.
[19] See generally Ryan Calo, Artificial Intelligence Policy: A Primer and Roadmap, 51 U.C. Davis L. Rev. 399 (2017).

cannot even identify whether their works were used, let alone assess the market harm that may have resulted.[20]

Disclosure requirements of this kind are already emerging in some jurisdictions. The EU AI Act imposes transparency obligations on providers of general-purpose AI models, requiring them to publish summaries of the content used for training.[21] The United States has no equivalent federal requirement as of 2026, but *Bartz* and its aftermath have created significant congressional pressure in this direction. Voluntary industry standards are insufficient: the competitive sensitivity of training data gives developers strong incentives not to disclose, and self-regulation has historically proven inadequate in markets where the interests of platform operators and content creators are structurally opposed.

C. Third-Party Auditing for Memorization

Dataset disclosure addresses the input side of the black box problem. Third-party auditing addresses the output side. Advances in interpretability research, the field concerned with understanding how neural networks process and represent information, have produced a suite of tools capable of detecting verbatim or near-verbatim memorization of training data in model outputs.[22] These tools are imperfect and their coverage is incomplete, but they are sufficient to provide a probabilistic assessment of whether a model's outputs bear a suspicious resemblance to specific training works.

A robust governance framework should require AI developers to commission and publish independent audits of their models' memorization properties before commercial deployment, with updates whenever the model is materially retrained. These audits should be conducted by accredited third parties with access to both the model's weights and a representative sample of the training corpus, and the methodology should be publicly documented to allow rights holders and courts to assess its reliability.

Auditing of this kind would not resolve the black box problem entirely because no currently available technique can provide a complete account of how a model's outputs relate to its

[20] For a proposal along similar lines, see Mark A. Lemley & Bryan Casey, Fair Learning, 104 Tex. L. Rev. 743 (2026).
[21] See EU Artificial Intelligence Act, Regulation (EU) 2024/1689, art. 13 (requiring 'transparency and provision of information to deployers' for high-risk AI systems).
[22] Vikas Hassija et al., Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence, 16 Cognitive Computation 45 (2024), https://doi.org/10.1007/s12559-023-10179-8.

training data. But it would provide courts with the kind of empirical foundation that the *Authors Guild* decision implicitly required: a verified, measurable account of what the model actually did with the copyrighted material, rather than a theoretical account of what machine learning pipelines are designed to do in general.[23]

D. Collective Licensing as a Market-Based Complement

The third element of the proposed framework is collective licensing. Even if dataset disclosure and auditing can make the fair use analysis more tractable, there will remain a class of uses that are clearly transformative, clearly lawful, and yet clearly harmful to rights holders in aggregate, not because any individual training run constitutes infringement, but because the cumulative effect of large-scale AI training is to depress the markets for human-authored creative works.[24]

Collective licensing models, modelled on the music licensing regimes administered by bodies such as ASCAP and BMI, offer a mechanism for compensating rights holders at scale without requiring case-by-case litigation. Under such a model, AI developers would contribute to a fund administered by a collecting society, which would distribute royalties to rights holders based on the degree to which their works were used in training.[28] The *Bartz* settlement, at approximately $3,000 per work, provides a rough baseline for what individual rights holders might expect to receive, though the adequacy of that figure remains contested.

Collective licensing would also address the structural power imbalance that makes individual litigation an inadequate remedy. The authors who brought *Bartz* were among the most commercially successful and legally sophisticated in their field; most writers whose works were ingested into AI training corpora lack the resources to pursue individual claims. A licensing regime would provide compensation to the broad class of rights holders affected by AI training, not merely those with the means to litigate.

V. CONCLUSION

Authors Guild, Inc. v. Google Inc. and *Bartz v. Anthropic PBC* are separated by a decade and by a fundamental shift in the technology they address. The Google Books project was large in scale but legible in operation: courts could inspect what had been done, measure its scope, and

[23] Li, supra note 5, at 1180.
[24] Li, supra note 5, at 1178.

evaluate its effects with reasonable confidence. Large language models are different in kind, not merely degree. They are opaque not only to courts and rights holders but, in important respects, to their own developers. The transformative use that *Bartz* accepted was a theoretical account of what these systems are designed to do, not an empirical finding about what this system actually did with these books.

That distinction matters because the fair use doctrine is, at its core, an empirical inquiry. It asks what was done, how much was taken, and what effect it had. When the answers to those questions are systematically inaccessible, courts are not applying the doctrine; they are deferring to the party with the more persuasive theory. In the AI copyright context, that party is almost invariably the technology developer, whose resources, technical expertise, and litigation capacity dwarf those of individual rights holders.

The governance framework proposed in Part IV, comprising mandatory dataset disclosure, third-party auditing, and collective licensing, is not a substitute for the fair use doctrine. It is the infrastructure that the doctrine needs to function. Without auditable training records, courts cannot apply the first and third factors meaningfully. Without memorization audits, they cannot assess the fourth. Without collective licensing, the gap between theoretical fair use and practical market harm will continue to grow.

Bartz v. Anthropic resolved a dispute. It did not resolve a problem. The black box remains. The question of whether AI truly transforms copyrighted material or merely repackages it at a level of abstraction that courts cannot presently inspect is not a question that litigation alone can answer. It requires the kind of institutional architecture that only legislation and industry regulation can provide. Until that architecture exists, the fair use doctrine will continue to function not as a principled balance between innovation and creators' rights, but as a de facto license for opacity.