READ, LEARN, INFRINGE? THE LEGAL PARADOX OF AI TRAINING DATA UNDER INDIAN COPYRIGHT LAW

Rutuja Bhand, LL.M. (IP and Tech Laws), Vishwakarma University, Pune

Rahi Ajabe-Alhat, Assistant Professor of Law, Vishwakarma University, Pune

ABSTRACT

Big AI models need tons of information to grasp how words, pictures, and trends work. From online sources, huge chunks get pulled without human help - books under copyright, news pieces, photos, lines of code, movies, artistic stuff. Here's the twist: these machines depend on duplication just to study, but making copies like that might break ownership rules. Places like the U.S., EU, and UK are shaping laws to handle this clash. Not so in India - no firm answer yet on if using protected content to train AIs counts as allowed.

Back when India made its Copyright Act in 1957, artificial intelligence wasn't even a thought. Because of that timeline gap, rules about copying, violations, or permitted use lack clarity on modern actions like pulling data online, building training sets, or teaching machines.

This study looks at parts 14, 51, 52, and 2(d) of India's Copyright Act, alongside court rulings including *Eastern Book Company v DB Modak* and *RG Anand v Deluxe Films*. Meanwhile, practices in the United States, Europe, and Britain are reviewed, especially how they handle limited usage, scanning material for patterns, and openness rules.

A fresh look at the findings shows India needs fairer rules for handling AI training data. These guidelines must support progress without ignoring those who create content.

Keywords: AI; Copyright; Fair Dealing; Data Mining; Copyright Violation; Indian Copyright Act; AI Regulation; Machine Learning; Generative; Intellectual Property Law.

Introduction

Out here, artificial intelligence shapes how people come up with ideas, work through them, then share what they find across digital spaces. Tools like ChatGPT from OpenAI, Google's Gemini, along with Stable Diffusion by Stability AI, produce writing, visuals, sound, software scripts, even imaginative pieces nearly indistinguishable from those made by humans. Behind the scenes, these models grow smarter after studying massive amounts of existing content pulled from both web sources and physical archives. Step by step, machines pick up skills by spotting trends, word arrangements, tonal nuances, connections hidden inside oceans of creations originally shaped by human minds.

What makes this legally tricky is how often training data includes things people own the rights to. Stuff like novels, academic studies, news stories, drawings, movie scripts, pictures, and computer programs gets duplicated and used when teaching AI.[1] Because of this, a strange situation shows up. Machines need access to protected material just to learn. At the same time, those who created it usually get full say over how it's shared or reused.

These days, India's copyright rules can't keep up with fast-moving technology. The original law started in 1957, long before computers could study millions of creative works in seconds.[2] Without clear updates, basic notions - such as what copying really means, whether use is fair, or who should be seen as an originator - feel shaky when AI dives into vast amounts of material. Artists feel uneasy, unsure if their work is still theirs. At the same time, developers face uncertainty trying to follow unclear lines. For creators and users alike, moving forward feels risky under outdated legal paths.

Here begins the courtroom stage, while lawsuits multiply across nations. A major clash appeared when The New York Times filed against OpenAI in America, saying its stories fed AI training without permission.[3] Close after, Getty Images moved on Stability AI, stating their image-driven models studied copyrighted visuals without green light.[4] One by one, these cases build a pattern - rapid innovation running headlong into older laws about who owns what.

[1] Pamela Samuelson, "Implications of Artificial Intelligence for Copyright Law," (2023) 21 Northwestern Journal of Technology and Intellectual Property 313.
[2] The Copyright Act, 1957 (India).
[3] *The New York Times Company v OpenAI Inc.*, ongoing litigation before U.S. courts.
[4] *Getty Images v Stability AI*, ongoing litigation before UK and U.S. courts.

For now, India has no solid laws or court rulings on if using data to train AI breaks copyright rules. It is still unclear if pulling information from websites counts as copying something. The legal weight of short-term duplicates made while machines learn sits without an answer.

Exceptions meant for fair use have not been clearly applied to artificial intelligence tasks.[4]

Research Questions

1. Does training an artificial intelligence with copyright-protected content constitute copyright infringement, legally speaking, in India?
2. Are there provisions of the Copyright Act of 1957 which can be applied in such cases of artificial intelligence training?
3. How have different countries handled the issue of AI training so far?
4. What kinds of reforms need to be made to establish a proper balance between innovations and rights of creators?

Research Objectives

1. To determine the legal status of AI training activities under Indian Copyright Law.
2. To evaluate the effectiveness of current laws for machine learning.
3. To discuss perspectives from across the globe regarding AI training.
4. To propose reforms which may prove beneficial for India.

Hypothesis

It has been hypothesised in this paper that the lack of sufficient provision in the Indian Copyright Act, 1957 to clearly delimit the legality of the use of AI training datasets has created considerable uncertainty for both content creators and developers of AI. The introduction of a legislative framework for text and data mining that allows for a greater degree of balance between innovation and the protection of IP is necessary to bring together innovation and

[4] Sections 14, 51 and 52, Copyright Act, 1957.

copyright protection.

Research Methodology

The study utilizes both a doctrinal and comparative methodology and utilizes both primary and secondary sources. Primary sources consist of the Copyright Act, 1957, judicial case law, international treaties, and statutes from other countries. Secondary sources include academic publications, policy papers, texts, and project reports on copyright law and artificial intelligence. Comparative analysis will be further compared to the legal framework of the United States, the European Union and the United Kingdom.

Understanding AI Training Data: A Technical and Legal Overview

Starting off, generative AI learns by going through a phase called training. It studies vast amounts of data to spot recurring shapes, links, or sequences in words, pictures, or similar inputs. While people grasp meaning through awareness, these machines don't experience comprehension like that. Their method relies on calculating likelihoods after seeing countless examples over time.[5]

A single word can spark a whole chain of thought when models study vast collections of written work, digging into how sentences flow and meanings connect across pages upon pages of texts. Paintings sit alongside snapshots in digital folders where programs trace lines, shapes, textures - learning not just what things look like but how they are arranged by human hands.

The data used for AI training includes:

- Books and literary works
- Stories from newspapers plus research papers
- Research papers
- Software code repositories
- Film scripts and subtitles

[5] Stuart Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach* (4th ed., Pearson 2021).

- Fanfiction and online forum content
- Music and audio recordings
- Photographs and digital artworks

Most artificial intelligence firms gather information using automatic methods like pulling pages from websites, scanning the web step by step, or gathering existing collections of digital material. Pulling pages means programs grab text and images from sites nonstop across thousands of locations. Step-by-step scans move through links in order, saving what they find into large pools meant for training models. Copies pile up behind the scenes, ready for machines to study patterns later.

Most rights holders have no idea their creations help train artificial intelligence systems. Instead of asking each artist directly, tech companies usually skip personal approval when pulling data together. Because these models require massive amounts of information, tracking down every single source just isn't practical in real-world terms.[6]

Out in the open, some cases show what happens when data sources get questioned. It turns out OpenAI admits pulling material from public corners of the web to shape its big language systems. Meanwhile, Google's Gemini finds itself under similar eyes - just like image-making tools built by Stability AI. One moment stood out: Stable Diffusion stirred noise once word spread about artwork, protected by copyright, being swept into learning pools without asking the creators.

Copying happens when AI learns, simply because data must move into machines to be studied. Though what comes out might look different, the act of pulling information in can still trigger legal rules around ownership. What matters is not just the result, but how it got there - storing even briefly could count as reproduction. Laws weren't built for this kind of movement, yet they apply once copies exist anywhere in the chain. Inside systems or not, duplication during learning stays a point of conflict.[7]

[6] WIPO, "Copyright and Artificial Intelligence," WIPO Conversation Paper (2020).
[7] Jane C Ginsburg & Luke Ali Budiardjo, "Authors and Machines," (2019) 34 Berkeley Technology Law Journal 343.

Indian Copyright Law and the Training Data Question

Who made something new gets special control under India's 1957 law on creative work. Yet nowhere in that rulebook is there a word about AI systems or how they learn. Since machines aren't named, people now ask if teaching them breaks those old rules. So judges and experts look closely at what's written - trying to see if feeding data to robots fits the idea of stealing someone else's creation.

Section 14: Exclusive Reproduction Rights

Ownership of creative content comes with special control over how it shows up physically.

Holding that right means deciding when a piece gets saved digitally, no matter the tech used.[8]

Since artificial intelligence needs vast amounts of data just to learn, copies often end up inside machines. Digital duplication happens early and often during setup, whether on servers or storage drives. Because of this, the law around making those copies suddenly matters a lot more than before.

So here's the big one. Could pulling data from websites count as copying under Section 14? That depends on what actually happens when a system grabs that info.

Maybe that's correct. Inside the system's learning phase, someone else's work ends up stored on machines after being duplicated. Processing follows, then it stays there - kept around so calculations can run later. Not for selling, just examining, yet duplication happened regardless.

Section 51: What Constitutes Infringement

Copyright gives certain rights only to the creator. That rule shows up clearly in Section 51.[9] Doing any of those special acts without permission breaks the law. A person might step into legal risk by using protected material carelessly. When companies copy creative work for AI training, questions follow. Approval missing means trouble may start. Using novels, songs, or articles without consent sits on thin ice. The law guards what belongs to others. Unauthorized reuse often counts as crossing a line. Machines learning from protected texts face scrutiny too.

[8] Section 14, Copyright Act, 1957.
[9] Section 51, Copyright Act, 1957.

Each unlicensed copy moves closer to violation.

Still, things get unclear since AI doesn't copy creative work like people usually do. Its aim isn't for someone to read or watch - it trains machines instead.

Section 52: Fair Dealing Exceptions

When it comes to copying, Section 52 allows some leeway for personal reasons, study, commenting on work, examining creative pieces, sharing news, along with select classroom activities.[10] While U.S. law gives wide room under "fair use," India's rules stick close to a fixed list, leaving little space beyond what is written.

It really comes down to one question: does using data to train AI count as study or acceptable use? Some tech firms might say they're just running automated tests, not copying creative work. Yet when those systems start making money, that argument begins losing strength.

Nowhere in India's laws is text mining clearly mentioned. Courts there usually keep a tight grip on what counts as fair use. Because of that gap, it remains unclear if Section 52 applies at all to how machines learn from data.

Section 2(d): Authorship

Who made what matters under Section 2(d), especially when it comes to books, art, music, or code.[11] Yet the rule leans hard on people being behind the work, leaving fuzzy gaps once machines start creating. When software learns from human pieces, then makes something new, the law stumbles. Clear lines blur fast - especially if no person directly shaped the output.

The Legislative Gap

It might be the biggest problem: India's copyright rules say nothing about using data to train artificial intelligence. Written long before machines could generate text or images, the law feels out of step now. Because it lacks clear guidance, judges could end up stretching old ideas into spaces they were never meant to cover.

[10] Section 52, Copyright Act, 1957.
[11] Section 2(d), Copyright Act, 1957.

Originality, Reproduction and Substantial Similarity

When it comes to AI training conflicts, India's legal approach around copyright holds useful ideas. Some of these have quietly shaped how such cases unfold over time.

Eastern Book Company vs DB Modak

Originality got a new twist in India when the Supreme Court weighed in on Eastern Book Company versus DB Modak.[12] Not long after, judges turned away from the old idea that effort alone earns copyright. A small spark of creativity became the key instead. Hard work matters less than a touch of fresh thinking. What counts now is not hours spent, but whether something shows even the faintest mark of invention.

When AI learns from creative material, questions about ownership start to surface. Because these systems study pieces made by people, the act of using them without permission can be problematic. The higher the level of imagination in a creation, yet the harder it becomes to justify its use unchecked. Learning involves repetition, though taking someone's unique output might cross a line.

RG Anand vs Deluxe Films

What matters most is whether the core form was taken, not just the concept behind it. Through RG Anand v Deluxe Films,[13] judges began using a test focused on deep likeness. Copying becomes legal issue only if what's borrowed goes beyond thoughts - into how those thoughts were shaped. The ruling made clear that protection lies in execution, never in inspiration alone.

Here’s a puzzle about how AI learns. Could it be grabbing specific styles, the kind protected by law? Or is it just pulling out general ideas, like shapes in data? What seems like imitation might actually be pattern spotting. Not every echo means copying.

Some who build AI might say systems only learn patterns across data, not exact reproductions. Still, others point out that when protected material shows up word for word, it suggests actual expression gets copied after all.

[12] *Eastern Book Company v D.B. Modak*, (2008) 1 SCC 1.
[13] *R.G. Anand v Deluxe Films*, (1978) 4 SCC 118.

Temporary Copies Doctrine

Here's a twist. Machines often make short-lived duplicates when learning from data. These fleeting versions pop up as part of normal operations. In certain places, the law allows such copies if they're essential to how tech works.[14]

For now, Indian law leaves it unclear if brief digital duplicates created while training machines count as copyright violation.

When Does Ingestion Become Infringement?

A tricky idea crops up when trying to pin down exactly where legal study ends and illegal duplication begins. Just opening a protected book brings no trouble for a person. But machines pull information into storage just by processing it, making the law stumble. That shift - from eyes on pages to circuits storing bits - changes everything without warning.

Comparative Analysis

United States

Floating inside America's legal framework, fair use stems from a clause tucked into the Copyright Act - specifically section 107.[15] This space allows bending without breaking laws.

Whether something counts as fair often depends on its goal, how it's made, how much is borrowed, what effect it has on sales.

Back in Authors Guild versus Google,[16] scanning books so people could search them was seen as fair game, given it changed their original form and helped society. That ruling pops up a lot when folks talk about teaching machines using texts.

Something shifts in training when imitation crosses into near-replica territory. Even so, legal fights - The New York Times versus OpenAI, for instance - reveal rising friction over where fairness ends.[17] Machines, some argue, only absorb structures. Yet a competing view insists

[14] Jane C Ginsburg & Luke Ali Budiardjo, "Authors and Machines," (2019) 34 Berkeley Technology Law Journal 343.
[15] Section 107, U.S. Copyright Act, 1976.
[16] *Authors Guild v Google Inc.*, 804 F.3d 202 (2d Cir. 2015).
[17] *The New York Times Company v OpenAI Inc.*, ongoing litigation before U.S. courts.

value flows directly from lifted labor. Disputes like those between Getty Images and Stability AI prove how blurry the boundaries still are.[18]

A twist in how courts handle cases shows adaptability, yet leaves room for doubt. Still, the system bends without breaking, even when outcomes feel unpredictable.

European Union

Now here's a twist - Europe carved out clear rules letting people copy digital content just to dig into patterns. Not everything fits though; only certain uses qualify, whether chasing knowledge or profit. Suddenly labs aren't the sole winners - businesses can join too, if they follow the lines drawn in 2019 laws.[19] Funny how one directive shifted who gets to explore massive piles of text.

Here's something key: the EU system lets creators step back if they don't want their work used in data scraping.

Out in the open now, details about training data must be shared by creators under the EU AI

Act 2024.[21] Since this rule landed, companies show how they meet standards - no exceptions. What's inside the models? That info can't stay hidden anymore. Following through means laying out steps taken to obey the law. Each developer handles it differently, yet disclosure remains fixed. Not optional. Clear records form part of the package. Rules push openness front and center. Hidden methods lose space here.

Out there among global efforts, Europe's approach stands as a broad effort to set rules for how artificial intelligence learns. While many regions watch and wait, this framework dives into what feeds machine learning. Though not perfect, it tries to cover more ground than most when shaping data standards. Where others skip details, this model pushes into specifics about sourcing and rights.

United Kingdom

A surprise twist in UK law allows reading machines to scan texts legally, thanks to Section

[18] *Getty Images v Stability AI*, ongoing litigation before UK and U.S. courts.
[19] Articles 3 and 4, Directive (EU) 2019/790 on Copyright in the Digital Single Market.
[21] European Union Artificial Intelligence Act, 2024.

29A of the Copyright, Designs and Patents Act.[20] Still, that freedom shrinks when profit enters the picture - only nonprofit study gets full protection.

Now beginning to look into changes around AI and copyright, the UK government moves as tech keeps shifting. With updates piling up, talks have started not because of pressure but timing. As new tools spread fast, officials choose this moment to rethink old rules. Driven less by urgency and more by necessity, the process opens quietly amid broader digital shifts.[21]

Apart from most places, the UK built rules into its legal system that recognize creations made by computers.[24] This sets it apart, giving shape to ideas formed without human hands.

Lessons for India

India can learn several lessons from comparative jurisdictions:

1. Clear laws beat guesswork in court. What's written matters more than what judges might think.

2. When rules allow limited use of digital material, new ideas may grow without harming original work. Though safeguards stay in place, progress often finds room to move.

3. Openness in rules makes people answer for actions.

4. Now and then, pulling away gives makers space to catch air during swift shifts. Power moves quietly when gear follows human rhythm instead of forcing it.

Still, India has to weigh its financial situation along with legal limits. Too loose rules could hurt local writers and publishing houses; too tight ones might slow progress in homegrown artificial intelligence work.

A fresh chance sits before India, one where a steady path can take shape around what matters most. This route might grow differently because local needs steer the way forward. Balance comes into play when choices align closely with real goals instead of outside models.

[20] Section 29A, Copyright, Designs and Patents Act, 1988 (UK).
[21] UK Intellectual Property Office, "AI and Intellectual Property Consultation," 2024.
[24] Section 9(3), Copyright, Designs and Patents Act, 1988 (UK).

Stakeholder Perspectives

Original Creators and Authors

Writers, painters, performers, reporters - each sometimes puzzled by where their work ends up today. Not always comfortable, the way things get shared. A photo spreads, a story twists, someone else takes it further. Creations drift beyond original intent. Faces appear on screens they never signed up for. Words quoted out of context, again. Even old sketches show up in places never imagined. The rhythm of making feels different now. Control slips, quietly.

Because machines learn from protected material, lots of makers say they get nothing when tech companies profit. Since artificial intelligence can produce similar pieces, people worry it might replace original work by humans.[22]

AI Developers

Still, tech firms worry tight copyright laws might slow progress down. Because machines learn by studying tons of material, builders say asking permission for each piece would break the whole process apart.

Still, they claim teaching machines helps people find knowledge faster while opening doors to learning. A boost shows up in how work gets done, along with fresh ways to invent things.

Gains appear across daily tasks plus new ideas taking shape.

Publishers

Caught between two pulls, publishers walk a tangled line. One direction tugs them toward guarding their content and earning from licenses. Yet, at the same time, certain ones form deals with artificial intelligence firms for profit.

Users and the Public

Out here, folks gain more access thanks to smart software helping them learn, write, explore ideas, and speak across languages. Still, people care about keeping original thinking alive -

[22] Pamela Samuelson, "Generative AI Meets Copyright," (2024) 72 Journal of the Copyright Society 1.

while pushing back when too much power gathers in a few hands.

Constitutional Balancing

Looking at things another way, Article 19 of the Indian Constitution brings certain principles into view.[23] Creative work falls within free speech, along with sharing ideas and pushing new technology forward. At the same time, those who make things have real claims - both financial and personal - to what they produce.

What makes it tough is finding space for clashing views in a system built on shared rule.

Legal Gaps and Proposed Reforms

Gap 1: No Explicit Provision Covering AI Training Data

Copyright rules in India say nothing about machine learning just yet. Training data sits in a gray area under current laws.

Reform Proposal

Introduce a statutory text and data mining exception specifically addressing AI training activities.

Gap 2: Fair Dealing is Too Narrow

Unlike U.S. fair use, Section 52 doesn't bend easily when it comes to computational tasks.[27] Its rigidity shows where automated methods need more room. While American rules adapt on a case-by-case basis, this part of the law stays fixed. That stiffness creates problems for datadriven work. Flexibility matters most when machines interpret content. Yet here, old structures remain unchanged. So digital analysis runs into limits sooner than expected.

Reform Proposal

Change Section 52 so that scanning texts and analyzing data is clearly allowed, but only within

[23] Article 19, Constitution of India, 1950.
[27] Section 52, Copyright Act, 1957.

set rules.

Gap 3: No Opt-Out Mechanism

Right now, Indian artists have no real way to stop their creations ending up in artificial intelligence systems. Though some voice concerns, few tools exist to block such use outright. Often, work gets pulled into training pools without warning or consent. Without clear rules, copying happens quietly behind the scenes. While debates drag on, material spreads across models regardless. Even original pieces risk vanishing into digital mimicry unnoticed.

Reform Proposal

Develop an opting out system that will enable rights holders to exclude works from AI training.

Gap 4: Absence of transparency

Most times, artificial intelligence firms can keep quiet about where they pull their training data. Yet transparency isn't forced by law in many regions. Because of that, users rarely learn what feeds machine learning models. Information gaps remain wide when rules don't demand disclosure. Without mandates, openness stays optional for tech developers.

Reform Proposal

Mandate transparency obligations requiring developers to disclose categories and origins of training data.

Compulsory Licensing Model

A fresh approach might take shape in India where AI builders chip in on a shared pot for artists. Instead of one-off deals this setup spreads pay through a central flow. Innovation keeps moving while those who make content get their share too.

Judicial Interpretation

For now, without new laws, judges in India could interpret current rules more thoughtfully when dealing with fresh conflicts. Seeing progress in tech matters just as much as defending the work behind original ideas.

Conclusion

Out of nowhere, machines started learning from creative work meant for humans. These systems soak up books, images, music - stuff protected by law - to build new outputs. Laws written long before algorithms could mimic style weren't built for such speed or volume. What once applied to people copying people now stumbles over code that learns in silence. Rules drafted for manual reproduction face a world where duplication happens inside silicon, unseen.

This work shows how India's rules on copyright do not clearly cover the use of data in training artificial intelligence. While laws about copying, violations, and fair use exist, they offer limited direction - raising unanswered questions for both artists and tech builders.

Looking closer, places like the EU and U.S. are already shaping laws around data scraping, usage rights, openness, and artist safeguards. Not so fast in India - rules there barely started forming.

A twist hides inside how laws handle AI learning - no single rule fixes it. Innovation might freeze if limits go too far; yet taking everything freely could weaken those who make stories, music, art.

Right now, India needs new laws fast - laws that balance invention with fairness for artists. Instead of picking sides between machines and human work, the real task is letting them grow side by side. Clear rules must show how creations are used. People should be able to say no when their work feeds AI systems. Payment structures ought to reflect true value, not vague promises. Progress means updating laws so neither tech nor art gets left behind. What comes next hinges on building trust, one honest rule at a time.

Bibliography

Books

1. Artificial Intelligence: A Modern Approach, 4th ed. (Pearson, 2021).

Journal Articles

1. Pamela Samuelson, "Implications of Artificial Intelligence for Copyright Law," (2023) 21 *Northwestern Journal of Technology and Intellectual Property* 313.

2. Jane C. Ginsburg & Luke Ali Budiardjo, "Authors and Machines," (2019) 34 *Berkeley Technology Law Journal* 343.

3. Pamela Samuelson, "Generative AI Meets Copyright," (2024) 72 *Journal of the Copyright Society* 1.

Cases

Indian Cases

1. Eastern Book Company v D.B. Modak, (2008) 1 SCC 1.

2. R.G. Anand v Deluxe Films, (1978) 4 SCC 118.

United States Cases

3. Authors Guild v Google Inc., 804 F.3d 202 (2d Cir. 2015).

4. The New York Times Company v OpenAI Inc..

United Kingdom and United States Litigation

5. Getty Images v Stability AI.

Legislation and Statutory Provisions

India

1. The Copyright Act, 1957.

2. Section 14, Copyright Act, 1957.

3. Sections 14, 51 and 52, Copyright Act, 1957.

4. Section 51, Copyright Act, 1957.

5. Section 52, Copyright Act, 1957.

6. Section 2(d), Copyright Act, 1957.

7. Constitution of India, Article 19.

United States

8. Section 107, U.S. Copyright Act, 1976.

European Union

9. Articles 3 and 4, Directive (EU) 2019/790 on Copyright in the Digital Single Market.

10. European Union Artificial Intelligence Act.

United Kingdom

11. Section 29A, Copyright, Designs and Patents Act, 1988 (UK).

12. Section 9(3), Copyright, Designs and Patents Act, 1988 (UK).

International and Institutional Materials

1. World Intellectual Property Organization, "Copyright and Artificial Intelligence," WIPO Conversation Paper (2020).

2. UK Intellectual Property Office, "AI and Intellectual Property Consultation," 2024.