THE RISE™

Exploring the Intersection of AI, Web Data Scraping, Copyright: Legal & Ethical Considerations

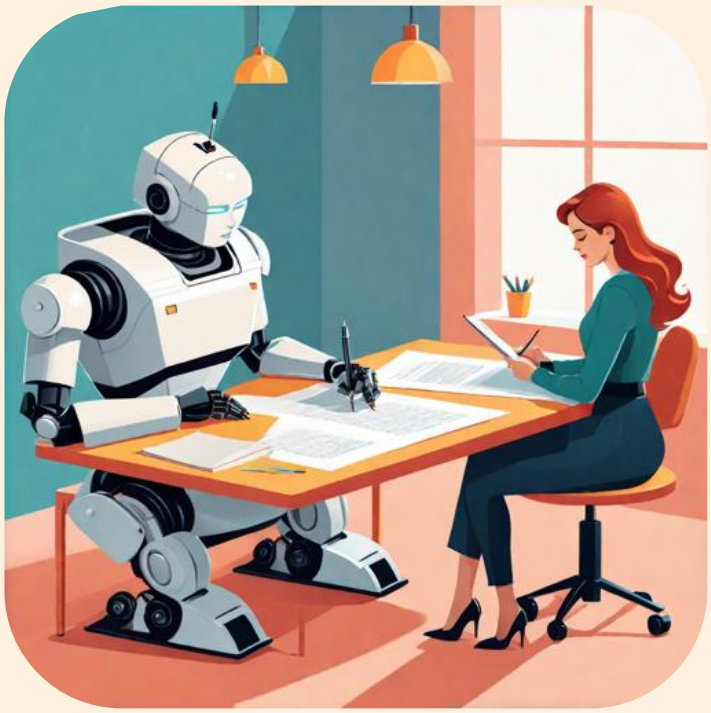
March 30, 2024

Loganayaki. P

Counsel at Svarniti Law Offices, New Delhi

Mitakshara Goyal

Co-Founder of Svarniti Law Offices, New Delhi



The fusion of web scraping and generative AI is revolutionizing data acquisition. Traditionally reliant on Python, web scraping now integrates AI for enhanced adaptability. Models like ChatGPT, DALL-E, and VALL-E demonstrate this shift, generating diverse outputs from online data. Leveraging vast training datasets from sources like Internet Archive and Wikipedia, these models adapt to the dynamic web landscape. Web scraping tools seamlessly supplement this process, ensuring continuous access to relevant information for effective AI training.

The Interplay of Web Scraping and Generative AI

The symbiotic relationship between generative AI and data acquisition relies heavily on web scraping, which systematically extracts information from diverse online sources. Integration of machine learning enhances adaptability to the dynamic web landscape, utilizing techniques like adaptive scraping and human-like browsing patterns to improve efficiency. However, ethical concerns and legal uncertainties arise as websites demand compensation and restrict access. The evolving legal landscape raises questions about liability for outputs generated by AI models trained on scraped data.

The symbiotic relationship between generative AI and data acquisition relies heavily on web scraping, which systematically extracts information from diverse online sources.

DNPA advocates for fair compensation for content used in AI model training, prioritizing transparency and ethical practices.

Clear boundaries between non-expressive and expressive AI uses are crucial, along with adherence to licensing requirements for responsible data usage.

DNPA's Advocacy for Copyright Protection in the Age of AI

The Digital News Publishers Association (DNPA), representing 17 major media publishers in India, confronts the complexities of AI models and their copyright implications. DNPA advocates for fair compensation for content used in AI model training, prioritizing transparency and ethical practices. Recent global

legal battles, such as *The New York Times* and Pulitzer-winning authors' lawsuit against *OpenAI* and *Microsoft*, underscore the need for robust measures. DNPA actively engages with Indian ministries to address industry concerns, seeking amendments to Information Technology Rules to protect publishers from potential copyright infringements by AI models. These efforts align with preparations for the forthcoming Digital India Act, expected to provide comprehensive regulatory frameworks for emerging technologies, including AI.

Copyright Challenges in AI Training

Legal experts acknowledge the void in the Copyright Act, 1957, concerning explicit provisions for AI training. In the absence of clear guidelines, the concept of 'originality' in AI-generated works may necessitate reconsideration. Indian courts may find themselves at the forefront of grappling with copyright infringement issues as AI increasingly relies on substantial datasets for training.

To navigate the legal landscape, Indian AI companies should emphasize originality in training datasets. They must carefully consider copyright notices, disclaimers, and ethical data feeding practices during training. While the possibility of authors receiving royalties from AI commercialization is uncertain, transparent practices and acknowledgment of original authors are crucial.

The New York Times vs. *OpenAI* and *Microsoft*

In a recent legal confrontation, *The New York Times* has initiated a lawsuit against *OpenAI* and *Microsoft*, shedding light on the expansive challenges in AI-related copyright issues. Central to this dispute is the claim that *OpenAI* utilized millions of articles without authorization to train its AI chatbots, transforming them into significant competitors to traditional news outlets. This legal action transcends a mere pursuit of monetary compensation, sparking broader discussions about the ethical dimensions of employing copyrighted materials in AI model training.

The New York Times' legal action goes beyond financial claims, seeking to address fundamental questions about the ethical use of copyrighted materials in AI development. The lawsuit suggests that *OpenAI's* AI chatbots have evolved into direct competitors, impacting the market and disrupting the established order in the media landscape.

OpenAI's Defense: Invoking the 'Fair Use' Doctrine

OpenAI vigorously defends itself by invoking the 'fair use' doctrine under the US copyright law. This argument asserts that training AI models with publicly available internet materials falls within the accepted bounds of fair use. While grounded in legal principles, this defense accentuates the ongoing ethical and legal debates surrounding AI applications and their interaction with copyrighted content.

The Judiciary's Response to Copyright Infringement by Generative AI

The judiciary's response to copyright infringement by Generative AI (GenAI) entities involves nuanced consideration of fair use exceptions. *OpenAI's* defense often aligns with the fair use doctrine, permitting the use of copyrighted material for purposes such as criticism, commentary, news reporting, teaching, scholarship, or research. Both in the United States and India, fair use assessment entails a multifaceted analysis, considering factors like the purpose and character of usage, substantiality of the portion used, effect on the market, and the nature of the copyrighted work.

Courts meticulously evaluate whether GenAI usage is transformative, contributing novelty or insights, or if it merely replicates the original work. This critical analysis helps determine the ethical and legal implications of AI practices and their alignment with established copyright doctrines.

The judiciary examines the portion used by AI entities like *OpenAI*, considering both quality and quantity, influencing infringement decisions. Courts assess AI's impact on the original work's market, determining substitution or complementarity. This evaluation guides fair use determinations, considering creativity and originality. The burden of proof in *The New York Times* case rests on demonstrating access and similarity, necessitating a clear link to the original work's market or value.

Copyright Laws and Data Scraping in India

Shifting focus to India, where data scraping implicates copyright laws, the Copyright Act of 1957 is a significant legal framework. Copyright protection extends to various works, and infringement occurs when the rights of the owner are exercised without proper authorization. Section 52 of the Act provides exceptions, including fair use for private use, criticism, review, or reporting of current events. Compliance with copyright law in data scraping hinges on adhering to these specified purposes.

Illustratively, in the case of *Eastern Book Company & Ors. v. DB Modak & Anr., 2008*, allegations of copyright infringement were contested. The court, after careful consideration, concluded that a requisite standard of creativity is necessary to claim copyright. While acknowledging the copyright in certain elements, it highlighted the need for creativity in the work for a valid copyright claim.

Legal Challenges in Data Scraping: Insights from the 2017 *Uber* Case

The 2017 *Uber* case underscored concerns regarding trade secret misappropriation and unfair advantage via data scraping. While the US has explicit laws governing trade secrets, India relies on equity or common law for breaches of confidence. India's legal stance on data scraping is uncertain, lacking direct references in existing laws. Stakeholders must assess terms of use and intellectual property implications, with adherence to best practices mitigating risks. Clear rules are crucial, especially considering the challenges in licensing datasets due to content diversity. Centralized databases could offer solutions for transparency and fair compensation.

Challenges in Fair Use for AI Training

Determining fair use exceptions for AI training, especially commercially, poses legal and ethical challenges. Clear boundaries between non-expressive and expressive AI uses are crucial, along with adherence to licensing requirements for responsible data usage. Protecting moral rights like paternity and integrity involves attributing human creators and considering intent for ethical AI practices.

Conclusion

The liability for AI-generated copyright infringement remains uncertain. Proposed measures suggest a comprehensive liability framework holding developers and users accountable. Mitigating legal risks entails defining responsibilities, complying with licensing terms, and conducting due diligence to promote ethical AI practices. The forthcoming Digital India Act and potential rule amendments signal a proactive approach to industry concerns, emphasizing the safeguarding of intellectual property rights in AI integration.