The Gap Between Ethical Expectations and AI Policy in Web Scraping

Philip Booth
Professor Doore
Colby College Department of Computer Science
December 5th 2025

Abstract:

Scaling laws have transformed the development of large language models by revealing that model performance increases predictably with the size of the training dataset. Once this relationship became clear, large AI companies accelerated their data acquisition efforts, creating a new landscape in which the demand for training data far exceeds what licensing alone can supply. During this period of rapid innovation propelled by scaling laws, the United States recognized that its regulatory environment might slow domestic AI progress relative to rival nations. In response, policymakers prioritized maintaining global AI competitiveness by relaxing oversight and allowing companies wide latitude in how they obtain data. These structural pressures have produced an expanding gap between what the law permits under doctrines like fair use and what stakeholders such as authors, technologists, and artists experience as ethically concerning. This paper examines the growing divide by arguing that the ethics of web scraping are inherently fraught, yet scraping has become structurally unavoidable for companies operating under the logic of scale. Using the Anthropic settlement as a case study, the analysis shows how legal permissibility obscures deeper issues of consent, data provenance, and creative ownership. The paper concludes that the only credible path toward an ethical future requires transparent documentation of training data and the continued role of class action litigation to confront and correct unlawful acquisition practices.

I. Introduction

Scaling laws show model performance improves when dataset size increases (Kaplan et al., 2020), and this discovery reshaped the priorities of modern AI development. Licensing currently cannot provide the volume or diversity required at this scale in a timely manner, so companies increasingly rely on web scraping because it is the main practical solution capable of producing corpora large enough for current models.

The U.S. has long been at the forefront of artificial intelligence innovation. As a result, the U.S. government has taken an almost utilitarian stance in favor of assisting American-based AI companies: "Imposing heavy-handed, intellectual-property-based restrictions on AI innovation will hamper the development of AI" (U.S. Copyright Office, 2024). Policymakers argue that limiting data access could slow progress, which strengthens the belief that regulatory slack is necessary for innovation. The sense of urgency reflects concerns about a widening gap between the GDPs of the EU and the U.S. and contributes to a climate often described as an AI arms race. In this context, data accumulation becomes both a technical necessity and a strategic obligation for companies that fear falling behind.

One of the key factors in rapid AI improvement has been the discovery of scaling laws by Jared Kaplan, who showed that a larger model generally equals a better model. The industry interpreted this to mean that competitive performance requires continuous growth of both models and datasets. This expectation intensified the rush to acquire as much training data as possible, which in turn expanded the distance between what is legally permitted and what many people experience as ethically troubling.

The Anthropic AI copyright class action litigation illustrates this divide. In its efforts to quickly scale its training efforts, Anthropic used torrent sites to upload pirated libraries of copyrighted books in addition to scanning books for training. Although the judge ruled that training on the scanned copyrighted books was protected under fair use, the judge took issue with the use of the pirate torrent sites, and the case revealed a deeper conflict created by scale itself. Competition spurred by unprecedented technology and the structural pressure to gather massive datasets encourages practices that remain both legally and ethically unresolved and push both current legal and ethical boundaries.

This paper argues that web scraping of protected content is ethically wrong yet structurally unavoidable, and that the only credible path toward an ethical future is to require rigorous transparency about data provenance and to rely on class action lawsuits to expose and correct unlawful acquisition practices.

II. Background

A. Why Web Scraping Became the Default for LLM Training

After the discovery of scaling laws, AI research shifted toward the emerging idea that model performance improves as data volume increases. Empirical work demonstrated that language-model performance increases smoothly as dataset size, model size, and compute are scaled up together, following predictable power-law trends (Kaplan et al., 2020). This insight launched a period of rapid technological advancement, where the new bulk of LLM training came in the form of data acquisition and upgrading computational resources. While licensed data acquisition is a powerful method for acquiring high-quality data in an ethical manner, this method alone cannot provide the volume or diversity required to continue scaling LLMs. Therefore, web scraping emerged as the dominant data acquisition method because it can harvest text at a scale, speed, and cost that licensed data sources cannot match.

B. How Data Scraping Works

1. *What is Data Scraping?*

Data Scraping is the automated extraction of information from websites or publicly available online platforms (Jayachandran & Arni, 2023). From a technical perspective, data extraction from the web occurs through three mechanisms: crawling, indexing, and API-based access. Crawling is the traditional approach for autonomously gathering large quantities of data. *Spiders* are set up at seed URLs and follow links to discover new pages. Indexing is the simple two-step process of acquiring website data in the form of HTML and extracting specific elements of the file. Most commonly, an HTTP request automatically visits a web page and downloads the HTML content. Then a parsing tool like the Python library Beautiful Soup reads through the HTML and extracts specific information by using identifiers such as tags, CSS selectors, and XPath expressions (Hage-Youssef and Cohen, 2025). API-based access remains an alternative to these aforementioned methods of web extraction. While websites are designed for human-facing interaction, APIs (Application Programming Interface) provide structured, standardized pathways for developers to access data in a controlled manner.

2. *The Modern Scraping Pipeline*

The following section outlines the web scraping pipeline from accessing websites to training LLMs with the scraped data.

a. *Crawling and Discovery*

While early crawling infrastructures relied on spiders to systematically follow hyperlinks from seed URLs, modern scraping pipelines incorporate AI-assisted strategies into crawling techniques. Traditional crawlers, using frameworks such as Scrapy, traverse through links to identify pages for extraction. AI-assisted crawlers can now navigate through dynamic websites and interpret shifting page layouts. As a

result, modern crawling methods extract more pages and are less prone to crashing.

b. Extraction

Once all pages are discovered, the pipeline moves onto extraction. There are two primary modes of extraction. Static extraction relies on HTTP requests to download a webpage's HTML, which is then parsed using a tool like Beautiful Soup to extract specific information. This method works well for sites that load content immediately (Hage-Youssef & Cohen, 2025). Conversely, dynamic extraction uses headless browsers (like Selenium or Playwright), which simulate user interactions to access gated web page content. For JavaScript-driven sites, this tool is essential in rendering dynamic content. Generative AI techniques have expanded extraction capabilities with vision-enabled LLMs that can extract images from websites that are not present in the HTML, as well as the elimination of the need for custom parsing codes for extraction. These further extractive properties of LLM-enhanced web scraping make the web both easier to scrape and remove limitations imposed by site owners.

c. Preprocessing Parsed HTML

The next step in the pipeline, content cleaning, transforms unstructured web data into dataset-ready text. This involves removing boilerplate elements, leaving only relevant text for extraction. Preprocessing also must ensure that meaningful text is passed in as training data, staying true to the familiar adage Garbage in, garbage out (Jayachandran & Arni, 2023). In the context of the pipeline, the preprocessing stage eliminates unnecessary or redundant material, leaving the final draft ready to be ingested by LLMs.

C. Data Acquisition Practices

AI companies have three primary mechanisms for data acquisition: licensing, bulk acquisition, and web scraping (Jayachandran & Arni, 2023). Crowdsourcing and Synthetic data generation also fall under this umbrella, but are bucketed more into the category of data generation, in which human intellectual property is of less concern. Each method offers a different type of sourcing in the broader ecosystem of training data.

1. Licensing

Licensing is essentially the authorized way to obtain data through a contract. AI companies enter into agreements with content owners that provide for the parameters of how the content/data can be accessed and used, and usually include a licensing fee for such use. A notable example is U.S. academic publisher Wiley, which reportedly made a $23 million deal with an undisclosed AI company for permanent access to all of its published articles (*Books+Publishing*, 2024). The question of whether a license is necessary in order to access copyrighted content is hotly debated and subject to many lawsuits between content owners and AI companies. AI companies generally claim that the training on copyrighted content constitutes a “fair use” under copyright law and that a license is therefore not necessary. Copyright owners, on the other hand, argue that making a copy of the data and using it for training the AI infringes their copyrights

and requires the express authorization of the copyright owner. Licensing provides the protocol for obtaining intellectual property in an ethical manner.

2. *Bulk acquisition*

Bulk acquisition involves downloading prepackaged datasets either online or offline. Downloading data does not require crawling infrastructure; otherwise it would qualify as scraping. While there are legal ways to acquire bulk data sets through bulk licenses, AI companies have also resorted to ethically questionable and potentially illegal means of bulk acquisition. The use of torrent sites to mass-download pirated books from known pirate sites like LibGen or Anna's Archive is such an example. Companies like Meta and Anthropic both face class-action lawsuits for attempting to circumvent copyright laws through the illegal mass acquisition of copyrighted books, and Anthropic settled its bulk piracy claim for $1.5 billion in the biggest copyright settlement in US history (U.S. District Court for the Northern District of California, 2025).

3. *Webscraping*

Web scraping is a powerful but minimally regulated mechanism for acquiring data. It is an extractive process that collects specific information from websites. Anything a human can access on the internet can be scraped and used to train AI models. Although web scraping is not designed solely for creating training datasets, scaling laws strongly incentivize its use (Kaplan et al., 2020). As a result, scraped data has become a key source of training material for large language models. The global web scraping industry is projected to surpass $9 billion by the end of 2025. Among all mechanisms for data acquisition, web scraping is the most efficient at collecting large quantities of data. Its lack of human intervention reduces manual error, and it is capable of locating high-quality data suitable for supervised learning models.

With the fast-paced innovation enabled by tools like web scraping, the practice creates a contentious legal and ethical environment. These issues will be further examined in the Ethical Concerns section, focusing specifically on infringement arising from the unauthorized acquisition of copyrighted materials.

III. Ethical Concerns

A. Transparency

Because data acquisition efforts for LLM training operate at a massive scale, AI companies obfuscate data collection practices, resulting in creators being unaware of how their works are utilized, obtained, and influencing model behavior. Schaul et al. (2023) concluded that "many companies do not document the contents of their training data—even internally —for fear of finding personal information about identifiable individuals, copyrighted material and other data grabbed without consent" (Shaul et al., 2023). Transparency in AI data scraping refers to the clarity and openness with which organizations

disclose their data collection, processing, and usage practices (Jayachandran & Arni, 2023). It is the responsibility of companies engaging in web scraping processes to inform website publishers if they are using their data. Transparency lies at the foundation of integrity, and without it, companies can hide a host of further ethical concerns like privacy, consent, and compensation. If users cannot understand how their work is being collected and how it is used, then they will be unable to confront the unlawful data practices in the first place.

Companies often promote themselves as ethically responsible to come across as moral actors. However, the misalignment of ethical beliefs of a company and their ethical practices is what is known as ethics shirking. Luciano Floridi defines digital ethics shirking in his book on AI ethics, the malpractice of doing increasingly less ethical work in a given context, the lower the return of that ethical work may be perceived to be" (Floridi, 2023). Ethics shirking in the US has become increasingly prevalent following Executive Order 15 U.S.C. 9401(3), cutting down on the legal repercussions of AI missteps in the name of innovation. This rhetorical transparency furthers the divide between what should happen and what actually occurs behind the veil of these powerful companies.

Anthropic is a key example of a company branded as an ethics-first AI company, whose brand name derives from the Greek word " anthrōpōs ", which roughly means "for the people". However, on September 5th, they settled for $1.5 billion following a class action lawsuit over the illegal use of downloading pirated books. It is only after companies have to reveal their internal data collection mechanism, in this case as part of a broader lawsuit over the legality of training, that illegal practices such as the use of torrenting copyrighted books from pirate sites become apparent. Without transparency, the authors lack agency and are given no voice in how their work is collected, used, or monetized.

B. <u>Data privacy</u>

By emulating a human browsing behavior, web scraping is engineered to acquire data in a way fundamentally misaligned with the intentions of website publishers. This design paradigm is extractive: it collects data regardless of the publisher's consent, repurposing data intended for human consumption instead to be ingested by a machine. As modern scraping techniques develop, so do their abilities to override non-consent. Because dynamic extraction tools can execute JavaScript and access otherwise hidden content, they defeat publisher attempts to limit the visibility of material, thereby overriding the contextual boundaries that govern privacy. Vision-based extraction allows scrapers to bypass HTML structure entirely, meaning that publishers cannot effectively prevent their visual content from being scraped. The result is a system that treats access restrictions as technical obstacles rather than understanding them as expressions of intent. Web scraping, therefore, raises privacy concerns not only

about the data that it scrapes but also about its mechanisms, explicitly designed and engineered to ignore publisher consent.

The practice of large-scale data scraping often collects more information than intended, resulting in the extraction of sensitive or private user data, which is embedded in web pages. Automated pipelines cannot distinguish between public context and private, hidden data. Vision-based extraction, for example, can scrape artwork from websites, whereas in an HTML file, the image would not be available in the same form. Additional examples of private data exposed to web scraping include metadata, personal identifiers, messages, location information, and unlisted images. Web scraping technology does not contextualize information extracted, either, collecting sensitive data indiscriminately from other text. Once a dataset is created, the owners who scraped the data can disseminate it to other parties. Fair data usage posits that data should be ethically sourced, respecting the privacy and intellectual property rights of creators and data subjects. A notable example is the MegaFace dataset, which included 3.5 million photos from Flickr, an online photo and video hosting platform, without the consent of the users who uploaded them. The data was redistributed widely to military and law enforcement agencies. Flickr users were neither informed nor asked for permission for their data to be repurposed in this way. The lack of transparency from MegaFace exploited both Flickr's commercial licensing structure and the trust between the platform and its user base, whose images were used to train technological systems without their consent. From this case, concerns about data provenance, privacy, and transparency emerge. While MegaFace was decommissioned following related legal challenges, no court required the deletion of already-scraped photos. This leaves unresolved questions about long-term data stewardship and the rights of the individuals whose images were included. Accidental capture becomes an inherent risk in data scraping, not only because of the unintended data collected, but also because these datasets are often redistributed outside the awareness of the individuals whose information comprises the dataset.

These privacy harms represent a fundamental gap between what is technologically possible and ethically permissible, a gap that current legal standards are not equipped to address.

C. Adherence to Legal Standards: Ethics Exposed in the Gap Between 'Fair Use' and Fair Treatment

While the question of whether training AI models on copyrighted material is copyright infringement or fair use is still unsettled, even the legal permissibility of training AI models on copyrighted material under fair use would not resolve the ethical issues that arise from such practices. In the United States, AI companies rely on the argument that training is a transformative practice because models do not reproduce the works but generate novel outputs. However, this reasoning obscures the reality that training

involves copying of expressive works. As Charlesworth argues in The Illusory Case of Fair Use, "content generated by an AI model based on a user prompt is inextricably bound up with the expressive exploitation of copied works," raising doubts about whether AI training can genuinely be considered transformative. The recent Anthropic class-action settlement illustrates this ambiguity. The company ultimately faced copyright infringement damages not for training on the books, but for acquiring the books through torrent sites. The judge found that the training itself was protected under fair use, but the means of obtaining the data were not. This split between legal training and illegal acquisition exposes the ethical challenge of providing recognition to creators whose work is being repurposed by AI.

This ambiguity is not limited to a single lawsuit. The difficulty in evaluating fair use for data scraping arises from a broader, unresolved landscape surrounding how training data is acquired and used. While certain groups seek payment for the use of their intellectual property in AI training, current doctrine does not require companies to obtain licenses for the massive amounts of scraped material involved. As a result, large-scale copying is often treated as permissible so long as a model's outputs are considered "transformative," a standard that focuses more on end results than on the extraction process itself. The law, therefore, concentrates on the harder copyright questions like copying, infringement, and transformation while overlooking the softer considerations of consent, provenance, privacy, and accountability. As the scale of scraping continues to increase, these unaddressed concerns widen the gap between what the fair-use doctrine evaluates and the ethical questions that arise from industrial-level data extraction.

D. How Licensing Barriers Enable the Shirking of Ethical Obligations

The difficulty in proposing refinements to fair use for data scraping stems from the complex and unclear landscape of determining the legality of scraping practices themselves. AI companies often justify their large-scale copying on the basis of how difficult it would be to obtain licenses for every individual piece of intellectual content, claiming that it is effectively impossible to negotiate with every rightful data owner whose work appears in their training sets (US Copyright Office Public Comments, 2023). They further argue that requiring statutory licensing agreements would slow or even stall domestic AI progress, a claim that frames scalability as a justification for bypassing traditional notions of consent. Yet this appeal to practical impossibility works less as a legal argument and more as a way of deflecting the ethical responsibility that normally accompanies the use of others' creative labor. The legal flexibility AI companies have received in the U.S., therefore, supports larger firms that can harness the extractive properties of data scraping to consolidate even more power. This dynamic is already visible in ongoing disputes: for example, The New York Times recently filed a lawsuit against Perplexity, alleging that the company's business model depends on scraping Times content to power its AI products (Guardian, 2025).

The difficulty of licensing, therefore, becomes less a technical obstacle and more a rhetorical strategy that normalizes extraction as the default mode of AI development. By positioning licensing as unworkable, companies shift attention away from the underlying question of whether creators should have any say at all in how their work enters these models.

IV. Case Study (Getty Images v. Stability)

The 2023 Getty Images v. Stability AI lawsuit represents a watershed moment in the ethical and legal considerations surrounding data scraping and AI. It is one of the first major copyright lawsuits challenging the scraping of creative property to train generative AI systems. Filed in both the U.K. and the U.S. in 2023, the lawsuit quickly became a test of whether existing intellectual property on the web can be legally used to train generative AI models. This lawsuit carries implications for ownership of creative works, data provenance, consent, and compensation.

The heart of the Getty Images case was the allegation that Stability AI, a company that develops an open-source platform for generating AI-produced content, scraped around 12 million copyrighted Getty Images photographs from its websites without consent and subsequently used them to train its AI models.

While the US case is still in its early stages, the UK case went to trial in the summer of 2025. Getty was forced to abandon many of its initial claims, including its primary claim for direct copyright infringement. There was no evidence that the training or development of Stability's diffusion model took place in the U.K. Getty argued that outputs of Stable Diffusion would sometimes contain the Getty watermark, indicating not only that the data had been scraped, but potentially that copyrighted images remained encoded in the model's learned representations. Getty concluded that the Stability AI model engages in a form of concealed reproduction that bypasses licensing and creator compensation while producing a commercial product built on their labor. Stability argued in response that imposing strict licensing requirements would inhibit innovation and concentrate power among firms with access to proprietary datasets.

The Getty lawsuit raises broader ethical questions surrounding data scraping. The central issue extends beyond the simple question of consent and compensation for creative labor. Getty's claim reveals a growing unease that, even if scraping were to constitute fair use, it does not respect the creative property rights of authors whose work fuels AI systems. The case illustrates the widening gap between existing copyright protections and the realities of large-scale automated data extraction. Although unresolved in the U.S., the U.K. branch of Getty Images v. Stability AI concluded with a finding that Stability AI's

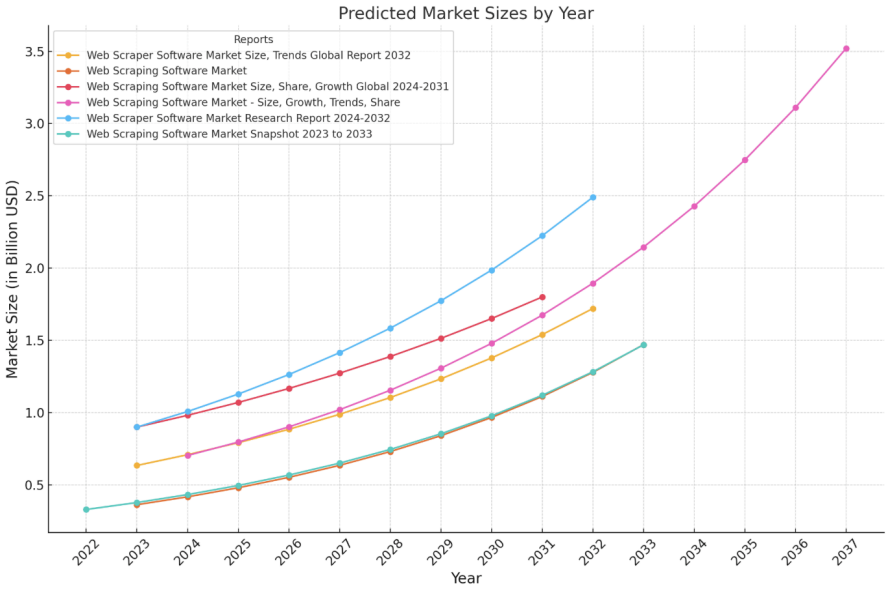
scraping practices complied with existing copyright standards and only took issue with the use of Getty's trademark on AI generated images.

A useful contrast to the Getty litigation comes from Anthropic's recent settlement over its use of pirated book copies obtained through torrent sites. Unlike the legal ambiguity surrounding Stability AI's data acquisition, Anthropic's situation involved clear copyright infringement. The company downloaded copyrighted books without authorization. In this case, court filings dismissed Anthropic's attempt to argue fair use, although the act of training on the data itself still constituted fair use.

Compared to Getty v. Stability, the Anthropic dispute underscores an important dimension of data-scraping ethics: provenance matters. Getty's claim hinges on whether scraping publicly visible content constitutes fair use. Anthropic's case demonstrates that when the underlying data is unambiguously pirated, technical transformation cannot justify fair use. This spectrum of cases reveals how courts are beginning to distinguish between lawful and unlawful data acquisition in the emerging AI landscape. Together, Getty and Anthropic illustrate that the ethics of AI development concern not only how models generate outputs but also the means by which the data is acquired.

V. Data

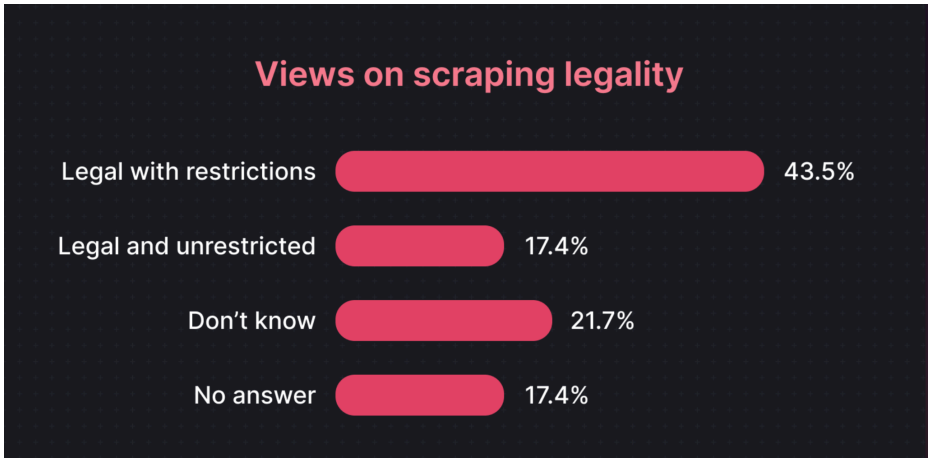
Pairing web scraping's rapid expected market growth over the next decade with the ambivalence surrounding the ethics of the tool itself makes web scraping deserving of scrutiny. The web scraping market is projected to grow at a CAGR above 11%, although estimates vary widely. ScrapeOps, a management tool for web scrapers, reported in its Market Report 2025 that this increase in demand can be heavily attributed to AI and machine learning, which require large-scale data to generate predictive analytics or train large language models. The report also notes that GDPR and CCPA compliance policies are reshaping the industry, with 86% of organizations increasing compliance budgets. The following graph is a prediction of the web scraping market size drawn from six different market growth reports.



<u>Figure 1</u>: Comparative projections of global web-scraping software market size across major industry reports, visualizing predicted growth trajectories from 2023-2037 (ScrapeOps Web Scraping Market Report 2025).

While clear variation exists in the initial 2023 valuation of the market, all of the reports predict strong CAGR growth, indicating a doubling or tripling of market size before 2033. As web scraping becomes a more critical tool across industries, allowing companies to leverage accessible online data, it prompts the need for revisions to the regulatory frameworks that govern the practice. Concerns around intellectual property rights, data privacy, and compliance with GDPR and CCPA highlight how regulation must evolve with the current landscape. Lawsuits filed against unethical scraping practices are becoming necessary battlegrounds for these concerns to take shape.

Since data scraping varies dramatically depending on the target website, the method of extraction, and the purpose of use, it must be scrutinized on a case-by-case basis. This makes these lawsuits central to shaping future ethical standards. Apify, a leader in third-party web scraping, published a report measuring public attitudes toward legal and ethical compliance in the industry.



Views on scraping legality among respondents.
Adapted from *State of web scraping 2024* (Apify, 2024).

These responses reveal the dissonance between policy and practice in web scraping. The majority of participants believe there should be stronger legal restrictions, yet current U.S. regulation permits the vast majority of scraping under fair use. The discrepancy between public sentiment and existing legal standards demonstrates the unresolved ethical tension surrounding web scraping and signals the need for more detailed analysis of how the practice should be governed.

VI. Ethical Framework Analysis and ACM code of ethics issues

A. Utilitarianism

Utilitarianism is an ethical framework intended to promote happiness and minimize pain in groups. As Mill more concretely puts it, "Utilitarianism is the creed which accepts as the foundation of the Greatest Happiness Principle holds that actions are right in proportion as they tend to promote happiness" (Mill, 2002). "Happiness in the right proportion" frames decisions as calculations that promote the greatest happiness to the greatest number of people.

As The executive order to deregulate certain AI policies for the sake of innovation is itself a utilitarian decision. By improving AI technologies that benefit society at scale, speeding innovation, even at the cost of weakening certain protections, can be justified under a utilitarian lens. Utilitarianism is also the only major ethical framework that can justify war, which explains why AI institutions temporarily reduce regulations in order to compete with other nations for AI dominance.

For governments that operate on a large scale, utilitarianism presents a compelling case because decisions can be calculated to benefit the greatest number of people. As AI companies begin operating at a scale similar to governmental institutions, OpenAI, for example, reports roughly 700 million weekly users, the same utilitarian reasoning becomes applicable to their decision-making.

At the micro level, data scraping embodies a utilitarian mindset. AI companies scrape data, including material that may break privacy norms or licensing laws, to deliver a product that will, in turn, benefit millions. Since utilitarianism justifies the means by the ends, the unethical practices that accompany AI innovation must be scrutinized so that harmful means are not indefinitely excused.

B. Deontology

While utilitarianism uses outcomes to justify actions, deontology places ethical weight on the rules and principles that govern those actions. These principles can be determined subjectively by the individual or institution employing the framework. However, once these principles are established, decisions are judged according to adherence to them, regardless of the outcome (Burton, 2023). Unlike utilitarianism, where the consequences are central and the means may be ambiguous, deontology prioritizes the means, even if the ultimate outcome cannot be controlled.

Since data scraping is a technology that largely operates through utilitarian logic, it inherently conflicts with deontological ethics. A deontologically oriented web scraper committed to strict ethical principles, for example, obtaining consent from every website or individual whose data is scraped, would find the process nearly impossible to execute at scale. The method would become infinitely slow and impractical.

Europe is not fully deontological in its approach to AI regulation, but its stricter legal frameworks and slower, more methodical processes reflect deontological influence. This leads to slower data acquisition but ensures stronger ethical protections throughout the process. This contrast highlights the tension between efficiency-driven innovation and principle-driven ethical restraint, an unresolved divide at the center of global AI governance.

C. Professional Code of Ethics

The ACM Code of Ethics asserts in section 1.3 that computing professionals have an obligation to be honest and trustworthy. As referenced earlier in the paper, many companies operating large-scale data extraction processes do not document the contents of their training data out of fear of finding personal information or other data obtained without consent. By opting out of disclosure, companies violate the

ACM Code by not “disclosing potential problems to appropriate parties” (Association for Computing Machinery, 2018).

In the hypothetical case that companies do disclose their data practices, they would expose themselves to a host of further scrutiny under the ACM Code, most notably section 1.6, which concerns respect for privacy. Web scraping practices are deliberately designed to bypass the non-consent of website publishers in order to extract as much information as possible. According to the ACM Code, “Computing professionals should only use personal information for legitimate ends and not violate rights of individuals and groups” (Association for Computing Machinery, 2018). Since gathering consent in web scraping for all sites visited is not a realistic solution, the ACM recommends that companies be transparent about the data they collect and how it is used so that affected individuals and groups may confront any non-consent on a case-by-case basis. By ACM standards, companies that do not implement this level of transparency are not adhering to their ethical obligations.

Lastly, according to section 1.5 of the ACM Code of Ethics, companies using web scraping processes to extract intellectual works must credit creators and recognize the value gained from using others’ work (Association for Computing Machinery, 2018). Because the current landscape of web extraction requires no agreement between the professional extracting information and the website publisher, this form of data acquisition remains largely unregulated. The recent Anthropic settlement demonstrates an agreement reached between the parties on how to recognize creative works and respect copyright, although Anthropic did not acquire the book data via scraping.

VII. Recommendations

In the same way that a single work scarcely alters a model’s behavior, individual creators have almost no power to confront AI companies that extract their data without consent. The first step toward restoring that power is to require transparency. Companies must meet rigorous documentation standards detailing the provenance of every training data source. Third-party auditability is essential to verify these disclosures and to ensure they cannot be selectively curated or altered. If the government mandates such documentation, the resulting paper trail would provide enforceable evidence in cases involving privacy violations, copyright infringement, or the misappropriation of creative labor.

An ETH Zurich article demonstrates that these requirements are not merely aspirational. Their open-source LLM, trained under the transparency obligations of the EU AI Act, achieved virtually no performance loss (ETH Zurich, 2025). This proves that ethical data governance is compatible with

state-of-the-art performance. Transparency is, therefore, not a hurdle for companies to overcome. Rather, it is through lifting the hood that unethical data practices become visible, and once visible, they can be challenged, regulated, and prevented.

When AI companies cross legal and ethical boundaries, class action lawsuits can force accountability and expose the widening dissonance between policy and regulation. Since extractive mechanisms like large-scale web scraping consolidate power in the hands of companies at the expense of individual recognition, class actions remain one of the only ways to impose external accountability and create incentives to uphold ethical standards. Edward Lee emphasizes that fair use "was never intended to apply categorically," meaning that copyright law requires a case-by-case evaluation of each act of copying rather than a blanket exemption for entire technologies like AI training (Lee, 2025). By passing soft ethical questions through class action litigation, independent creatives can establish a relationship with AI companies and draw boundaries around the ways in which their works are used.

The bittersweet outcome of Anthropic's class action settlement indicates that one-time compensation for intellectual property is fundamentally unfair. Despite the $1.5 billion pool, authors received only token payments (less than $3,000 per book after legal fees), an amount that neither reflects the market value of their creative work nor the ongoing contribution those works make once absorbed into a training corpus. The settlement exposes a deeper structural problem: AI models extract continuous value from datasets, while creators receive compensation only once, if at all. This imbalance mirrors the broader "coercive calculus" of scaling laws, where creative labor becomes a raw material for technological progress rather than a protected form of human expression.

A sustainable governance model for AI should therefore shift from one-time compensation toward systems that protect creators over the entire lifecycle of their work. Establishing a standardized licensing or royalty framework for text, similar to structures that compensate musicians when their songs are streamed, would allow authors to retain an ongoing stake in the commercial value their writing generates once used in training data. This process would transform creators from invisible inputs into recognized contributors. By combining transparency requirements, documentation standards, and mechanisms for continuous compensation, policymakers can build an AI environment that respects creative labor while preserving the pace of technological innovation.

VIII. Conclusion

The legal debates surrounding fair use and data scraping reveal just how unprepared existing frameworks are for the pace of modern AI development. What counts as permissible under current doctrine often fails

to address the deeper ethical concerns raised by the mechanisms of acquiring and using data. As lawsuits like the Anthropic case show, policy and ethics can diverge, leaving creators without recognition or control over their work. At the same time, transformative-use rulings continue to exacerbate the extractive dynamics that shape who benefits from AI progress. Only by enforcing transparent provenance, strengthening consent requirements, and preserving the role of litigation can we narrow the widening gap between legally permissible scraping and ethically responsible innovation.

IX. Bibliography

Apify. (2024). *State of web scraping 2024*. https://blog.apify.com/state-of-web-scraping/

Association for Computing Machinery. (2018). ACM code of ethics and professional conduct. https://www.acm.org/code-of-ethics

Bartz, A. (2025, October 3). *The thriller writer who took on a tech giant*. *The New York Times*. https://www.nytimes.com/2025/10/03/books/review/andrea-bartz-anthropic-lawsuit.html

Books+Publishing. (2024, September 4). *Wiley expects to make US$44 million from AI partnership, authors unable to opt out*. https://www.booksandpublishing.com.au/articles/2024/09/04/258068/wiley-expects-to-make-us44-million-from-ai-partnership-authors-unable-to-opt-out/

Burton, Emanuelle, Judy Goldsmith, Nicholas Mattei, Cory Siler, and Sara-Jo Swiatek. *Computing and Technology Ethics: Engaging through science fiction*. Cambridge, MA: The MIT Press, 2023.

Charlesworth, J. C. (2025). Generative AI's illusory case for fair use. *Vanderbilt Journal of Entertainment & Technology Law, 27*, 323–356. https://ssrn.com/abstract=4924997

Cohen, M. C., & Hage-Youssef, E. (2025, July 16). *Generative AI for data scraping*. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5353923

ETH Zurich. (2025, July 9). *A language model built for the public good*. https://ethz.ch/en/news-and-events/eth-news/news/2025/07/a-language-model-built-for-the-public-good.html

Fieldfisher. (n.d.). *Data scraping: Considering the privacy issues*. Retrieved December 5, 2025, from https://www.fieldfisher.com/en/services/privacy-security-and-information/privacy-security-and-information-law-blog/data-scraping-considering-the-privacy-issues

Floridi, L. (2023). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford University Press.

Jayachandran, Jayasankar and Arni, Vijay, Traversing the Ethical Landscape of Data Scraping for AI (December 8, 2023). https://ssrn.com/abstract=4666354 or http://dx.doi.org/10.2139/ssrn.4666354

Lee, E. (2025). *Fair use and the origin of AI training*. *Houston Law Review* (forthcoming). Santa Clara University Legal Studies Research Paper No. 5253011. https://ssrn.com/abstract=5253011

Metz, C. (2025, September 5). *Anthropic agrees to pay $1.5 billion to settle lawsuit with book authors*. *The New York Times*. https://www.nytimes.com/2025/09/05/technology/anthropic-settlement-copyright-ai.html

Prolific. (n.d.). *AI, data scraping, ethics and data quality challenges*. Retrieved December 5, 2025, from https://www.prolific.com/resources/ai-data-scraping-ethics-and-data-quality-challenges

Publishers Association. (2025, September 5). *FAQ: Anthropic settlement case* [PDF]. https://publishers.org/wp-content/uploads/2025/09/FAQS-Anthropic-September-5-2025-1.pdf

Samuelson, P. (2009). Unbundling fair uses. *Fordham Law Review, 77*(5), 2537–2621. https://ir.lawnet.fordham.edu/flr/vol77/iss5/16

ScrapingBee. (n.d.). *How to scrape all text from a website for LLM AI training*. Retrieved December 5, 2025, from https://www.scrapingbee.com/blog/how-to-scrape-all-text-from-a-website-for-llm-ai-training/

U.S. Copyright Office. (2024). *Copyright and artificial intelligence: Part 3, generative AI training* [Pre-publication report]. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf

United States District Court for the Northern District of California. (2025). *Notice of $1.5 billion proposed class action settlement between authors & publishers and Anthropic PBC: Bartz v. Anthropic PBC, No. 3:24-cv-05417-WHA*. https://assets-us-01.kc-usercontent.com/1eeb16db-4934-006e-40a6-38fa91285ebb/7ce4b15b-2240-43a8-8b40-e92625b5f25b/ANT%20-%20Long-Form%20Notice%2011.21.2025_10_43_PM.pdf

Zirpoli, C. T. (2023, February 24). *Generative artificial intelligence and copyright law* (CRS Legal Sidebar LSB10922). Congressional Research Service. https://digitalcommons.unl.edu/scholcom/243