INTERNATIONAL JOURNAL FOR LEGAL RESEARCH AND ANALYSIS

Open Access, Refereed Journal Multi Disciplinary
Peer Reviewed

www.ijlra.com

ISSN: 2582-6433

DISCLAIMER

No part of this publication may be reproduced, stored, transmitted, or distributed in any form or by any means, whether electronic, mechanical, photocopying, recording, or otherwise, without prior written permission of the Managing Editor of the *International Journal for Legal Research & Analysis (IJLRA).*

The views, opinions, interpretations, and conclusions expressed in the articles published in this journal are solely those of the respective authors. They do not necessarily reflect the views of the Editorial Board, Editors, Reviewers, Advisors, or the Publisher of IJLRA.

Although every reasonable effort has been made to ensure the accuracy, authenticity, and proper citation of the content published in this journal, neither the Editorial Board nor IJLRA shall be held liable or responsible, in any manner whatsoever, for any loss, damage, or consequence arising from the use, reliance upon, or interpretation of the information contained in this publication.

The content published herein is intended solely for academic and informational purposes and shall not be construed as legal advice or professional opinion.

ISSN: 2582-6433

ABOUT US

The *International Journal for Legal Research & Analysis (IJLRA)* (ISSN: 2582-6433) is a peer-reviewed, academic, online journal published on a monthly basis. The journal aims to provide a comprehensive and interactive platform for the publication of original and high-quality legal research.

IJLRA publishes Short Articles, Long Articles, Research Papers, Case Comments, Book Reviews, Essays, and interdisciplinary studies in the field of law and allied disciplines. The journal seeks to promote critical analysis and informed discourse on contemporary legal, social, and policy issues.

The primary objective of IJLRA is to enhance academic engagement and scholarly dialogue among law students, researchers, academicians, legal professionals, and members of the Bar and Bench. The journal endeavours to establish itself as a credible and widely cited academic publication through the publication of original, well-researched, and analytically sound contributions.

IJLRA welcomes submissions from all branches of law, provided the work is original, unpublished, and submitted in accordance with the prescribed submission guidelines. All manuscripts are subject to a rigorous peer-review process to ensure academic quality, originality, and relevance.

Through its publications, the *International Journal for Legal Research & Analysis* aspires to contribute meaningfully to legal scholarship and the development of law as an instrument of justice and social progress.

ISSN: 2582-6433

PUBLICATION ETHICS, COPYRIGHT & AUTHOR RESPONSIBILITY STATEMENT

The *International Journal for Legal Research and Analysis (IJLRA)* is committed to upholding the highest standards of publication ethics and academic integrity. All manuscripts submitted to the journal must be original, unpublished, and free from plagiarism, data fabrication, falsification, or any form of unethical research or publication practice. Authors are solely responsible for the accuracy, originality, legality, and ethical compliance of their work and must ensure that all sources are properly cited and that necessary permissions for any third-party copyrighted material have been duly obtained prior to submission. Copyright in all published articles vests with IJLRA, unless otherwise expressly stated, and authors grant the journal the irrevocable right to publish, reproduce, distribute, and archive their work in print and electronic formats. The views and opinions expressed in the articles are those of the authors alone and do not reflect the views of the Editors, Editorial Board, Reviewers, or Publisher. IJLRA shall not be liable for any loss, damage, claim, or legal consequence arising from the use, reliance upon, or interpretation of the content published. By submitting a manuscript, the author(s) agree to fully indemnify and hold harmless the journal, its Editor-in-Chief, Editors, Editorial Board, Reviewers, Advisors, Publisher, and Management against any claims, liabilities, or legal proceedings arising out of plagiarism, copyright infringement, defamation, breach of confidentiality, or violation of third-party rights. The journal reserves the absolute right to reject, withdraw, retract, or remove any manuscript or published article in case of ethical or legal violations, without incurring any liability.

ISSN: 2582-6433

HARVESTING ADVANTAGE: ANTITRUST IMPLICATIONS AND UNFAIR COMPETITIVE PRACTICES ARISING FROM AI DRIVEN DATA SCRAPING

AUTHORED BY - HARSHIT MADAAN & NIKHIL KUMAR JHA

ABSTRACT

The rapid expansion of artificial intelligence has transformed data into a critical competitive resource, with data scraping emerging as one of the primary methods through which large datasets are collected for training AI systems. This paper examines the legal and competition-law implications of automated data scraping, particularly in the context of AI-driven markets where control over data can significantly influence market power and innovation. While data scraping enables technological development and facilitates the functioning of large language models, it simultaneously raises concerns relating to intellectual property, privacy, and unfair competitive practices.

The Paper analyses how existing legal frameworks have predominantly addressed scraping disputes through copyright and computer-access laws, often overlooking the broader competition concerns that arise from data monopolisation. Through a comparative examination of foreign jurisprudence, including cases such as Van Buren v. United States, hiQ Labs v. LinkedIn, and ongoing litigation involving AI developers, the paper highlights how courts have increasingly narrowed the application of computer-access statutes while leaving unresolved the structural competition issues created by large-scale data aggregation.

Thereby, the paper evaluates India's regulatory framework, focusing on the Competition Act, 2002. It argues that while statutes like Copyright Act address aspects of unauthorised access and intellectual-property protection, they fail to provide a coherent mechanism for dealing with the competitive consequences of large-scale data collection and control. Comparative analysis with the European Union's Digital Markets Act and the United States' antitrust approach demonstrates alternative regulatory strategies for managing data-driven market power.

ISSN: 2582-6433

INTRODUCTION: DISSECTING DATA SCRAPING FOR STRUCTURED DATA EXTRACTION

The intricate storytelling of a Studio Ghibli film or the unexpected coherence of a well-curated Reddit thread reflects a kind of intelligence that today's AI strives to replicate, not through consciousness, but through patterns drawn from zillions of data reservoirs. From predictive models to creative generation, Artificial Intelligence (hereinafter referred to as AI) is becoming more attuned to human nuance. Yet, as seen in the Reddit–Anthropic lawsuit, the value of this data and who controls it are now at the centre of a heated debate. Data may power AI, but ownership still writes the rules.

AI systems are varied in their scope and functionality. One such category is Generative AI, which refers to an AI model with the capability of learning to generate content such as images, text, and other media with similar properties to its training data, as well as systems built on such models. These platforms, classified as large language models (LLMs), are trained on vast corpora of text data. Through this training, they internalise linguistic patterns and contextual relationships. When presented with a user prompt, the model generates a response by predicting the most contextually appropriate sequence of words, producing output that closely resembles natural human language.

AI systems, such as Anthropic's Claude, are trained using three principal categories of data: publicly available information, datasets supplied by third parties through business agreements, and data contributed by users or workers on associated platforms.[1] The first and third sources have become subjects of ongoing legal debate, public data raises issues related to proprietary rights, while user-generated data brings forth significant concerns regarding privacy and informed consent.

The automated process of collecting data from websites without authorisation is precisely what is referred to as data scraping, a practice whose legality continues to be widely debated. It involves a three–tier process including extracting raw, unstructured information, followed by filtering and curation to produce datasets suitable for AI training. The process involves digital tools such as web crawlers to extract data from targeted websites, which is then used by bots

[1] Shepard Goldfein and James Keyte, 'Big Data, Web 'Scraping' and Competition Law: The Debate Continues' (2017) 258(49) The New york Law Journal.

ISSN: 2582-6433

to train LLMs.[2] Yet, at the heart of this seemingly technical process lies an undeniable reality: the data was extracted without permission.

In HiQ Labs v. LinkedIn,[3] the Ninth Circuit Court of Appeals observed that both parties claimed to serve the public interest by advocating for the free flow of information on the internet. Supporters of data scraping argued for open access, the opposing side warned that permitting such access could enable malicious actors to infiltrate their systems and compromise server security. As a result, they contended, protective measures would become necessary,[4] potentially restricting access to data that was once public and ultimately undermining the broader public interest. Therefore, the Project strives to fine-tune the two positions, whereby consideration of Antitrust Laws, alongside commercial viability, also needs to be addressed.

JURISPRUDENTIAL STANCE OF COURTS ON LEGALITY OF DATA SCRAPING

Foreign Jurisprudence and the Competition Law Lacuna: A Greater Threat; Problematic Tilt Toward Copyright Remedies

Unauthorized access to computer systems remains the doctrinal threshold in data-scraping litigation across jurisdictions, yet the dominant legal response has disproportionately centred on copyright and computer-access statutes rather than on competition law. This copyright-centric orientation is problematic: it treats the conflict as a property or private-law dispute, thereby obscuring the more systemic risks that data monopolisation poses to market structure, entry and consumer welfare. In consequence, the comparative jurisprudence demonstrates that the absence of robust competition-law engagement, not merely gaps in copyright, presents the more acute, long-term threat to contestability in data-driven markets.

Leading authorities illustrate this dynamic. In "Van Buren v. United States (2021)[5] the U.S. Supreme Court constrained the reach of computer-access offences by adopting a "gates-up / gates-down" test: liability under statutes like the CFAA turns on circumvention of

[2] Audrey Pope, 'NYT v. Open AI: The Times's About-Face, Harvard Law Review' (Harvard Law Review June 20, 2025) https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/ accessed 29 August 2025.
[3] *HiQ Labs, Inc., v Linkedin Corporation*, No. 17-16783 (2022).
[4]Anthropic, 'how do you use personal data in model training' (Anthropic) https://privacy.anthropic.com/en/articles/10023555-how-do-you-use-personal-data-in-model-training 2 September, 2025.
[5] *Van Buren v United States* No. 19–783 (2020).

ISSN: 2582-6433

technological access barriers, not on improper use of legitimately accessed data. Applying that principle, the Ninth Circuit in HiQ Labs v. LinkedIn,[6] held that scraping publicly accessible LinkedIn profiles did not contravene the CFAA where no authentication gates were bypassed. While these rulings appropriately limit criminal and quasi-criminal enforcement, they also shift the burden of addressing exclusionary data conduct away from criminal law and toward civil regulatory regimes, notably competition law, which, in many jurisdictions, has not yet developed adequate doctrine for data access disputes.

Concurrently, high-profile civil litigation against AI vendors has concentrated on copyright infringement, Digital Millennium Copyright Act (DMCA) theories, and related private-law claims. In The New York Times Co. v. OpenAI,[7] and Microsoft plaintiffs allege large-scale ingestion of copyrighted articles and removal of copyright-management information (CMI); OpenAI has asserted fair-use defences and jurisdictional objections. Raw Story Media v. OpenAI,[8] demonstrates courts' insistence on particularised proof of concrete harm, a requirement that often narrows copyright plaintiffs' remedies where market injury is speculative. Cases such as Kadrey v. Meta Platforms,[9] Andersen v. Stability AI,[10] Getty Images v. Stability AI,[11] and CanLII v. Caseway AI,[12] show that copyright and DMCA claims may proceed where plaintiffs plausibly allege systematic ingestion and substantial similarity between model outputs and protected works. Thomson Reuters v. Ross Intelligence[13] further exemplifies the decisive role of market-effect analysis in fair-use adjudication: copying of proprietary headnotes for a competing legal-research product was held not to be fair use because it was commercial, non-transformative and likely to impair the plaintiff's market.

These copyright-focused remedies, while important to right-holders, are inherently limited in addressing the competition problem. Copyright litigation tends to (i) require proof of protectability under originality tests, (ii) hinge on substantial-similarity or CMI allegations, and (iii) deliver remedies that are primarily exclusionary (injunctions or damages) without structured consideration of market-wide consequences or calibrated access remedies (for

[6] *HiQ Labs, Inc., v Linkedin Corporation* No. 17-16783 (2022).
[7] *The New York Times Company v Microsoft Corporation, Open AI Inc.* 23-cv-11195 (SHS) (2025).
[8] *Raw Story Media v OpenAI* 24 Civ.01514 (2024).
[9] *Richard Kadrey, et al. v Meta Platforms, Inc.* 23-cv-03417-VC (2023).
[10] *Andersen v Stability AI* 3:23-cv-00201 (N.D. Cal. Jan 13, 2023).
[11] *Getty Images v Stability AI* [2025] EWHC 38 (Ch).
[12] *Canadian Legal Information Institute v Clearway management Ltd. Alistair Vigier as Caseway AI, Caseway AI Legal Limited* VLC-S-S-247574 (2024).
[13] *Thomson Reuters v Ross Intelligence* 1:20-cv-613-SB (2025).

ISSN: 2582-6433

example, time-limited licensing, non-discriminatory API obligations, or behavioural commitments).[14] As a result, courts and regulators that rely predominantly on copyright frameworks risk leaving unaddressed the broader phenomenon of data entrenchment: dominant platforms can still leverage public or quasi-proprietary datasets to foreclose rivals, even where copyright claims fail or are inapplicable.

In short, foreign jurisprudence reveals a twofold lesson: criminal-access statutes have been narrowed (placing fewer disputes within CFAA-style enforcement), and copyright litigation, though sometimes successful, is an imperfect tool for mitigating exclusionary data practices. The more urgent institutional lacuna is therefore competition law: absent doctrinal and enforcement upgrades addressing data as a potential essential input and regulating refusals to deal or discriminatory access, markets remain vulnerable to durable concentration that copyright litigation alone cannot remedy.

Indian Precedential Analysis of Data Scraping vis-à-vis Copyright Emphasis Versus Competition-Law Deficit

Indian jurisprudence has not yet issued a settled rule on automated data scraping; however, the Supreme Court's copyright jurisprudence on compilations provides the principal domestic lens through which many scraping disputes are presently viewed. That copyright-first orientation, while legally significant, risks sidelining competition-law analysis and thereby failing to address structural competitive harms arising from concentrated control of data.

In Eastern Book Company v. D.B. Modak,[15] the Supreme Court rejected the sweat-of-the-brow" doctrine and established that compilations attract protection only where they reflect the exercise of skill and judgment. The Court adopted a "skill and judgment" threshold (informed by CCH Canadian Ltd. v. Law Society of Upper Canada),[16] holding that merely mechanical or clerical editorial acts do not generate copyright. This test is consequential in the AI context: whether a news feed, database or curated repository is protectable determines the availability of copyright remedies against scraping and, by extension, whether denial of access may be

[14] Office of the Australian Information Commissioner, 'Guidance on privacy and developing and training generative AI models' (Office of Australian Information Commissioner) https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-developing-and-training-generative-ai-models (accessed 30 August 2025).
[15] *Eastern Book Company v D.B. Modak* (2008) 1 SCC 1.
[16] *Canadian Legal Information Institute v Clearway management Ltd. Alistair Vigier as Caseway AI, Caseway AI Legal Limited* VLC-S-S-247574 (2024).

ISSN: 2582-6433

legally justified on IP grounds.

The pending ANI v. OpenAI,[17] litigation before the Delhi High Court will test these principles. ANI asserts that OpenAI ingested proprietary news content and that model outputs have caused reputational and commercial injury. Indian courts adjudicating such claims will be required to resolve two core enquiries: (1) whether the disputed material satisfies the Eastern Book Company standard and thus merits copyright protection against ingestion and storage for training; and (2) whether the defendant's conduct falls outside statutory fair-dealing exceptions, Section 52,[18] once the four-factor analysis of purpose, nature, amount and market effect is applied.

Crucially, however, this copyright-centred pathway is not an adequate substitute for competition-law engagement. Even where compilations are protectable, exclusive control over datasets can produce foreclosure effects in adjacent markets, effects that copyright relief (damages or injunctions) may mitigate only partially and in ways that do not recalibrate market structure. Conversely, where compilations fail the originality threshold, competition authorities may have stronger grounds to intervene; yet CCI doctrine and guidance on data access, refusal to deal, and essential-facility matters remain underdeveloped.

Consequently, Indian adjudication faces a two-pronged evidentiary burden if plaintiffs pursue copyright remedies: (i) they must show their works meet the required skill-and-judgment standard; and (ii) they must adduce convincing proof of material ingestion or downstream reproduction causally linked to quantifiable market or reputational harm. But even successful copyright claims leave unresolved macro-competitive questions: Should a dominant platform that accumulates non-protected but commercially indispensable data be permitted to deny access? If so, on what terms? Those questions require competition-law instruments, policy frameworks, doctrinal calibrations, and remedies, rather than exclusive reliance on copyright doctrine.

Therefore, the current doctrinal emphasis on copyright protection, while necessary to vindicate creators' rights, is insufficient and potentially misleading as the principal response to data-scraping controversies. Without parallel and purposive development of competition law

[17] *Ani Media (P) Ltd. v Open AI Inc* 2024 SCC OnLine Del 8120.
[18] The Copyright Act, 1957 (14 of 1957) s 52.

ISSN: 2582-6433

principles, recognising data as a potential essential input, articulating refusal-to-deal standards, and equipping the CCI with technical and remedial tools, India risks permitting data-driven market entrenchment that neither copyright nor contract law can adequately remedy.

REGULATORY STANCE VIS-À-VIS INDIAN AND FOREIGN ANTI-TRUST FRAMEWORK

Indian Legislative Framework

Under the Indian legislative framework, data scraping currently falls within the purview of two statutes: the Information Technology Act, 2000[19] and the Copyright Act, 1957.[20] Section 43(b) of the IT Act prescribes penalties for "unauthorised access" involving the extraction or downloading of content from a computer system or network. However, a critical limitation arises in its application: the provision does not extend to the scraping of information that is already available in the public domain.

This lacuna is reinforced by "Rule 3 of the Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011,[21] which expressly excludes publicly accessible information from the ambit of "sensitive personal data or information" (SPDI). As a result, the unauthorised harvesting of publicly available data, such as open website content, may escape liability under the current statutory framework.

The ambiguity surrounding the scope of "unauthorised access" has been clarified in comparative jurisprudence. Notably, the U.S. Supreme Court in Van Buren v. United States,[22] adopted a narrow construction, holding that liability attaches only when technical barriers or authentication measures are circumvented. This interpretation, if harmonised with Indian law, would highlight the inadequacy of the present framework and underscore the necessity of refining statutory definitions to address the complexities of data scraping.

The Competition Act, 2002:

The Competition Act, 2002,[23] contains no express reference to "data" or "scraping," and its

[19] The Information Technology Act, 2000 (21 of 2000).
[20] The Copyright Act, 1957 (14 of 1957).
[21] Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, G.S.R. 313, r 3.
[22] *Van Buren v United States* No. 19–783 (2020).
[23] The Competition Act, 2002 (12 of 2003).

ISSN: 2582-6433

provisions are applied on the basis of economic effect and market context. Section 3 of the Act prohibits anti-competitive agreements.[24] In principle, any collusive arrangement among firms to restrict the availability or use of data – for example, a cartel whereby competitors agree to share or fix prices using scraped data, could violate Section 3(1) (prohibition of cartels and other anti-competitive agreements).[25] Similarly, vertical arrangements that effectively bar rivals from accessing a platform's data might fall under the vertical restraints in Section 3(4).[26] For instance, Section 3(4) bans vertical supply or distribution agreements that include a "refusal to deal" with certain classes of persons where this causes an appreciable adverse effect on competition (AAEC).

Thus, if a dominant platform imposed contractual terms forbidding trading partners from enabling scraping by competitors, that could be scrutinized as an agreement to foreclose access under Section 3. In practice, however, most scraping issues arise from unilateral conduct rather than formal agreements, so Section 3 would only apply if the conduct were anchored in a collusive or exclusive-dealing contract. The principal focus lies under Section 4, which prohibits abuse of dominance.

Section 4(1) and (2) bar a dominant firm from engaging in exclusionary or exploitative conduct. Notably, Section 4(2)(c),[27] outlaws denial of market access, and Section 4(2)(e),[28] forbids unfair or discriminatory conditions, both of which can encompass refusal to share data or interfaces. Although the Act does not list "data" explicitly, Indian precedent treats proprietary information or software interfaces as essential inputs. For example, in MCX Stock Exchange Ltd. v. National Stock Exchange, the CCI found that a dominant stock exchange's refusal to share its trading API ("APIC") with a rival was an abuse of dominance. The Commission characterized the NSE's proprietary trading software as an "essential facility" and held that denying access to it was a denial of market.

By analogy, a dominant digital platform that bars competitors from programmatic access (e.g. via an API or by actively blocking web crawlers) could be deemed to have denied an essential input under Section 4(2)(c). In Schott Glass India, the Supreme Court emphasized that such a

[24] The Competition Act, 2002 (12 of 2003) s 3.
[25] The Competition Act, 2002 (12 of 2003) s 3(1).
[26] The Competition Act, 2002 (12 of 2003) s 3(4).
[27] The Competition Act, 2002 (12 of 2003) s 4(2)(c).
[28] The Competition Act, 2002 (12 of 2003) s 4(2)(e).

ISSN: 2582-6433

refusal must be judged by its competitive effect: a total denial of supply "falls within Section 4(2)(c)" only if it causes an AAEC. Thus, Indian law requires an effects-based inquiry: even if scraping is restricted, liability under Section 4 depends on whether the conduct appreciably harms competition.

The "restrictive" form of scraping is therefore conceptually covered: a refusal by a dominant platform to allow scraping (e.g. by refusing to provide data access or by discriminating among users of its service) can be challenged as a refusal to deal or discriminatory practice under Section 4. Although no reported Indian case yet involves web scraping per se, the NSE decision suggests that denial of data interfaces is treated akin to denial of market access. By contrast, the "enabling" form – namely, a firm itself undertaking mass scraping – is more complex. Unilateral scraping does not fit neatly into any enumerated abuse, but if a dominant firm's data extraction were shown to foreclose rivals or facilitate collusion, it could be caught as an exclusionary abuse. For instance, if a dominant news aggregator scraped all relevant content and thereby stifled competing news sites, that might be seen as harming the competitive process. Indeed, the CCI has recently expressed concern over Google's "zero-click" news snippets, holding that scraping and displaying publishers' content allows Google to earn revenue while depriving publishers of traffic.

Although that investigation is ongoing, it indicates that mass scraping by a powerful firm could be viewed as injurious to competition under Section 4 if it reinforces dominance or denies rivals' fair participation. Therefore, while the Competition Act is neutral on data rights, its broad prohibitions can sweep in scraping-related conduct depending on economic effect. Section 3 would catch any collusive or restrictive agreement involving data-sharing or no-scrape clauses. Section 4,[29] would apply to unilateral conduct by a dominant firm that either denies access to data (analogous to refusal to deal or uses data in an exclusionary way. Key authorities (e.g. Schott Glass and NSE) confirm that refusal to deal must have an appreciable adverse effect on competition. In the absence of explicit statutory language, Indian enforcement will turn on evidence of foreclosure, discrimination, or other anticompetitive economic impact caused by data scraping practices.

[29] The Competition Act, 2002 (12 of 2003) s 4.

ISSN: 2582-6433

The Digital Competition Bill, 2024:

India currently has **no law explicitly defining or banning "data scraping"**. The Information Technology Act, 2000 generally forbids unauthorised access to computer systems (Section 43), but courts have not applied it to routine extraction of publicly posted information. Under India's new Digital Personal Data Protection Act 2023, "publicly available" personal data is largely exempt from the law.In other words, data that an individual has put in the public domain can be collected and processed without consent.This reflects a liberal stance: collection of open or non-sensitive data is essentially unregulated, whereas personal data (even if public) technically requires consent.

In practice, India relies on existing frameworks rather than a specialized scraping regime. The proposed *Digital Competition Act"* (a draft bill recommended in early 2024) would create ex ante rules for dominant "Systemically Significant Digital Enterprises" (SSDEs), For example, the draft forbids SSDEs from using non-public data of third-party businesses to compete against them, and generally mirrors concepts from the EU's Digital Markets Act.

However, it does **not** enact a distinct anti-scraping provision; its data-related rules focus on privacy, fairness, and interoperability (e.g. barring mandatory bundling of services). Beyond legislation, private websites often deter scraping contractually (through terms of use) or technically (via robots.txt directives), but these measures have never been tested in India's courts. In sum, India's regulatory approach treats data scraping largely as a by-product of privacy and competition law, not as a standalone offence.

Judicial and CCI Precedents: Indian courts have addressed scraping only in **intellectual-property cases**, not as a competition issue per se. In *OLX BV & Ors. v. Padawan Ltd*, the High Court granted a permanent injunction against a UK-based site that had "lifted" OLX's listings by automated means. The court found that copying OLX's copyrighted content and trademarked elements (via scraping) amounted to infringement and passing off, and restrained the defendant from any further data extraction. Aside from OLX, no major Indian court or the CCI has ruled directly on scraping.

However, Indian competition regulators have explicitly recognised **data accumulation itself as a source of market power**. The Competition Commission of India (CCI) noted in a recent report that "access to data can… represent a form of competitive advantage," since online firms

ISSN: 2582-6433

amass vast consumer information to fuel personalised services. Reflecting this view, the CCI in 2023 penalised Meta (Facebook) (₹213 crore) for abusing its dominance by forcing WhatsApp users to share data with other Meta companies. Although that case arose under competition law (Sections 4(2)(c) and 4(2)(e) of the Competition Act)[30] and privacy policy changes, it shows the CCI's willingness to treat unfair data practices as anti-competitive. No Indian decision has held a scraper liable solely for accessing public data: unlike in the US, where *hiQ Labs v. LinkedIn,*[31] allowed scraping of public profiles, Indian courts have not had occasion to decide whether Section 43 or other laws reach routine public-data scraping.

Assessment of India's Approach: India's current stance is a mix of liberalism and laissez-faire. On one hand, exempting public-domain data from both IT Act penalties and data protection obligations means that most scraping of open information is effectively **permitted**. This can foster innovation (e.g. AI training on scraped web data) and keeps entry barriers low for new players. Indeed, by not imposing broad anti-scraping rules, the law avoids chilling beneficial uses of non-sensitive data. On the other hand, the regime has notable gaps. There are **no mandatory data-sharing obligations** on dominant platforms (in contrast to e.g. EU proposals), so large incumbents can hoard data.

Enforcement is **fragmented**: multiple bodies (CCI, the forthcoming Data Protection Authority, sectoral regulators like RBI/SEBI, etc.) have overlapping claims on digital data policy. Policymakers themselves have acknowledged potential jurisdictional overlaps among regulators and called for ironing them out. In practice, websites rely on self-help (privacy policies, contracts, robots.txt) to bar scraping, but no Indian court has ruled on whether violating these barriers yields liability. In short, India depends on general competition law and contract principles to check anti-competitive scraping, which keeps the framework flexible but risks **weak enforcement** and uncertainty for users and firms alike.

Comparative Perspectives – EU and US: Other jurisdictions take varied approaches. In the **European Union**, data scraping is governed primarily by privacy and competition laws. The Digital Markets Act,[32] a novel ex ante competition regulation for gatekeepers – explicitly prohibits designated platforms from combining or "cross-using" personal data across different

[30] The Competition Act, 2002 (12 of 2003) s 4(2)(c).
[31] *HiQ Labs, Inc., v Linkedin Corporation*, No. 17-16783 (2022).
[32] European Union Digital Markets Act, 2022 (2022/1925).

ISSN: 2582-6433

services without explicit, separate user consent. In effect, a gatekeeper cannot scrape and merge business user data or consumer data from one service to benefit another without the user's clear agreement. The DMA also mandates interoperability and data portability for consumers, to prevent lock-in. Earlier, the EU Court of Justice had held in *Ryanair Ltd. v. P.R. Aviation BV*,[33] that even scraping non-personal data (flight schedules) from a competitor's site was prohibited without permission. These rules reflect the EU's view that protecting competition and privacy requires firm controls on how data is gathered and used by powerful platforms. (Separately, the EU's 2022 Data Act encourages data sharing from IoT devices, but it focuses on machine-to-machine data rather than open-web scraping.)

In the **United States**, there is no specific anti-scraping law. Scraping of publicly available data is generally allowed under existing statutes. Notably, in "*hiQ Labs, Inc. v. LinkedIn Corp.*,[34] the court held that scraping LinkedIn's public profiles did **not** violate the Computer Fraud and Abuse Act (CFAA). The CFAA only bars access to systems "without authorization" when access is technically or legally restricted; the court ruled that freely accessible LinkedIn pages were not covered, and allowed hiQ to continue harvesting them. Thus, in the US, the primary check on scraping is contract or copyright law (or state "trespass" statutes), not new regulation. US antitrust authorities have investigated data issues (for example, challenging tech mergers on data grounds) but have not yet enacted a Digital Markets-style law. The FTC and state attorneys general may address abusive data practices on a case-by-case basis. In summary, the EU uses a **preventive, rules-based** model (ex ante DMA obligations, strong privacy law) to curb abusive scraping, whereas the US relies on **ex post enforcement** (antitrust actions, technical-law interpretations) and has generally treated public-data scraping as lawful.

India could consider a mix of legal clarifications and new rules to address scraping in a balanced way by **clarifying permissible scraping of public data; DMA-style consent requirements.** Following the EU example, India could mandate that large digital firms obtain **separate, informed consent** before combining data from distinct sources (especially business user vs. consumer data). This was recommended by the German Bundeskartellamt and is reflected in the EU DMA; an Indian equivalent (perhaps in the draft Digital Competition Act) could prohibit gatekeepers from intermixing datasets without user approval. Similarly, the draft Digital Competition Bill already bans SSDEs from using non-public business-user data to

[33] European Union Court of Justice, *Ryanair Ltd. v. P.R. Aviation BV* (Ryanair Ltd.) (2015) C-30/14.
[34] *HiQ Labs, Inc., v Linkedin Corporation*, No. 17-16783 (2022).

ISSN: 2582-6433

compete with them – this could be extended into a broader obligation requiring explicit consent or data-sharing mandates across core services.

Regulatory coordination. Overlaps between the CCI, the data protection authority, and other regulators are a concern. The government should clarify each body's mandate on data issues. For instance, setting up a formal **inter-agency council or protocol** for digital data policy (similar to EU's Single Market oversight) could ensure that competition and privacy objectives are aligned, preventing duplicative or contradictory actions.

Promote interoperability and portability. The DPDP Act's data portability provisions are a step in the right direction. Further reforms could require dominant platforms to implement open APIs or data access interfaces, as the EU does for payment data. India might also explore a "Consumer Data Right" regime, like Australia, for certain sectors to ensure data can move with the user,[35] which would indirectly discourage anti-competitive scraping practices by making data lock-in harder. **Technical and best-practice guidelines.** The government could issue guidelines on "responsible scraping" (e.g. mandating rate limits, respect for privacy when processing data) and on adherence to robots.txt, giving clarity to businesses on compliant data collection. It could also encourage norms like attribution or data-sharing consortia, drawing from global "data trust" or "collaboration" models.

In essence, reform in India should strike a balance: **permit benign scraping of public data"** (to foster innovation and AI) while equipping regulators to check abusive or privacy-intrusive data harvesting by dominant firms. Learning from the EU's ex ante consent and interoperability rules and the US's focus on outcomes, India could amend its laws and improve coordination so that competition policy and privacy law together guard against data-driven monopolies.

Comparative Perspective: EU and US

In the EU, competition law likewise does not mention "scraping," but Article 102 TFEU,[36] has long recognized that denying competitors access to essential inputs can be abusive. The CJEU held in Magill and IMS Health that a refusal to license intellectual property or information can

[35] Deng J, 'Antitrust Regulation on Refusal of Data Scraping and Sharing: The Application of the Essential Facility Doctrine' https://www.cresse.info/wp-content/uploads/2020/02/2019_ps4_pa1_Antitrust-Regulation.pdf accessed 2 September 2025

[36] Treaty on Functioning of the European Union (adopted 25 March 1957, entered into force 1 January 1958) (1958) (TFEU) art 102.

ISSN: 2582-6433

be an abuse if the asset is indispensable and refusal eliminates competition on a downstream market. Recent EU enforcement (e.g. Google Shopping) and the new Digital Markets Act (DMA) go further: the DMA explicitly obliges designated "gatekeepers" to share certain data with business users.

For example, Article 6(11) of the DMA,[37] requires large search engines to provide search data (rankings, queries, clicks) to competing search engines. In this way the EU has imposed ex ante data-access rules on digital platforms. EU authorities have signaled that abusive data scraping (or denial of scraped data) will be judged by traditional competition principles – focused on consumer harm and market foreclosure – and by tailored digital rules. By contrast, US antitrust has traditionally been wary of imposing data-sharing mandates. Under Sections 1 and 2 of the Sherman Act,[38] unilateral refusals to deal are generally lawful unless a monopolist's change in course lacks any procompetitive justification.

In Trinko the US Supreme Court cautioned that forcing a firm to share its "source of advantage" is disfavored, because it may deter investment and create "central planning" problems. Therefore, US courts generally decline to extend an "essential facilities" duty beyond narrow precedents like Aspen Skiing (in which a monopolist had abruptly stopped cooperative conduct with a rival). Nonetheless, US regulators and courts are increasingly attentive to data access. The Ninth Circuit in HiQ v. LinkedIn,[39] observed that allowing a dominant platform to selectively exclude competitors from public data "may well be considered unfair competition" However, that case was decided under state-law unfair competition claims, not Sherman Act per se. In practice, US enforcement focuses on merger reviews and Section 5,[40] unfair methods of competition, rather than using Section 2 to compel data access. In sum, while EU competition law and regulatory reform actively address data scraping (through both antitrust cases and the DMA) US law remains more deferential, treating data access as a property right absent clear evidence of consumer harm.

[37] European Union Digital Markets Act, 2022 (2022/1925) art 6(11).
[38] Sherman Act, 1890 (15 U.S.C. Note 1) s 2.
[39] *HiQ Labs, Inc., v Linkedin Corporation* No. 17-16783 (2022).
[40] Sherman Act, 1890 (15 U.S.C. Note 1) s 5.

ISSN: 2582-6433

REFORMING THE COMPETITION ACT, 2002: ADDRESSING THE LACUNAE ON DATA SCRAPING AND AI-DRIVEN MARKET POWER

Several areas of Indian law require urgent reform to address emerging technological complexities; among these, competition law is particularly in need of statutory reinforcement. The rise of data scraping and AI-driven business models has converted datasets into competitively decisive inputs. Yet, the Competition Act, 2002 does not presently articulate principled rules for when and how control over datasets translates into exclusionary market power. This legislative silence creates a systemic risk: disputes that in reality concern market structure and access are treated primarily as private copyright or contract conflicts, thereby leaving broader harms to contestability, innovation and consumer welfare unremedied.

A central doctrinal tool for addressing these harms is the Essential Facilities Doctrine. Under the framework articulated in MCI Communications Corp. v. AT&T Co.,[41] and applied in comparable jurisdictions, liability for refusal to deal turns on four elements: a defendant's control over an indispensable facility; the inability of the claimant to reasonably duplicate that facility; the defendant's denial of access; and the feasibility of providing access on reasonable terms. Translating this reasoning into the Indian context requires that the Competition Act or authoritative CCI guidance recognise datasets (including aggregated, curated, or user-generated databases) as potential essential inputs where they are non-replicable and indispensable for effective competition in a downstream market.

International authorities already illustrate why competition law must play a leading role. U.S. and U.K. case law, notably hiQ Labs v. LinkedIn,[42] and Authenticom v. CDK Global,[43] reveals that denying access to publicly available or platform-controlled data can have exclusionary consequences even when criminal-access statutes prove inapplicable. The Google-Yelp episode further demonstrates how a dominant platform may both appropriate rivals' downstream content and coerce acceptance of unfavourable terms, thereby foreclosing competition in adjacent services. These examples underscore that privacy or proprietary arguments may be invoked strategically to mask anticompetitive exclusion; consequently, Indian law cannot leave such disputes to bilateral contract or copyright litigation alone.[44]

[41] *MCI Communications Corp. v. AT&T Co.* 512 U.S. 218 (1994).
[42] *HiQ Labs, Inc., v Linkedin Corporation* No. 17-16783 (2022).
[43] *Authenticom v CDK Global* 17-2540 & 17-2541 (2017).
[44] Ioannis Drivas, 'Liability for Data Scraping Prohibitions under the Refusal to Deal Doctrine: An Incremental Step toward More Robust Sherman Act Enforcement'

ISSN: 2582-6433

Against this background, the Competition Act should be revised, either by amendment or through binding CCI guidelines, to set out a data-sensitive refusal-to-deal test. Such a test should evaluate: (i) whether a data-holding firm exercises control over a dataset that is practically non-duplicable; (ii) whether a rival's inability to access the data materially impairs its ability to compete; (iii) whether access can be granted on reasonable, non-discriminatory and commercially viable terms without undermining legitimate privacy or IP interests; and (iv) whether less intrusive remedies (time-limited access, remunerated licences, interoperability commitments) can restore competitive parity. The test must explicitly require that privacy and copyright considerations be subjected to proportionality review and not be allowed to operate as per se justifications for foreclosure.

Merger control must also be recalibrated to address data consolidation. The notification and appraisal criteria under Sections 5, 6 should require specific analysis of data overlaps and the risk of creating "data moats" that entrench dominance even where turnover thresholds appear modest. Where acquisitions concentrate unique, non-replicable datasets, the CCI should be empowered to impose tailored conditions, ranging from access undertakings to structural remedies, to preserve ex ante contestability.

Finally, effective enforcement will demand institutional upgrades. The CCI should develop in-house technical expertise or formal mechanisms for secondees to assess data provenance, replication costs, and the technical feasibility of proposed access remedies. Procedural safeguards for confidentiality and data protection must be incorporated into investigational powers and discovery rules. Coordination protocols with the Copyright Office and data-protection authorities should be established so that IP, privacy and competition concerns are adjudicated in a coherent, complementary manner rather than in isolated silos.

Therefore, the Competition Act must be proactively modernised to address the distinctive features of data-driven markets: recognition of data as a potential essential input, a calibrated refusal-to-deal doctrine, enhanced merger scrutiny for data consolidation, proportionate access-oriented remedies, and strengthened technical and procedural capacity for enforcement. Without such reforms, India will remain exposed to a regime in which copyright and contract law are over-relied upon to resolve fundamentally competition-centric problems, an outcome

https://lawreview.uchicago.edu/sites/default/files/Drivas_CMT_Final.pdf accessed 3 September 2025.

ISSN: 2582-6433

that risks entrenching market power and impeding the very innovation the law should promote.

CONCLUSION AND SUGGESTIONS

This project has explored how India's competition law framework addresses the challenges of data scraping, particularly in an era shaped by artificial intelligence and platform dominance. Data scraping, the automated collection of web information, sits at the intersection of intellectual property, privacy, and competition law. Internationally, copyright law has been used to address scraping disputes, but it often proves inadequate because originality thresholds are high and remedies are limited to injunctions or damages. As a result, many jurisdictions have turned to competition law, recognising data itself as a source of market power.

In India, courts have not yet directly ruled on scraping, and most disputes have been resolved under copyright principles. The Supreme Court's approach in *Eastern Book Company v. D.B. Modak,*[45] raised the bar for database protection, leaving many scraped compilations outside copyright's reach. Pending disputes like the ANI v. OpenAI,[46] case highlight this tension, where fair use may defeat claims of infringement even if substantial scraping has occurred. Yet the broader competition problem remains: a firm may lawfully scrape large volumes of data, then leverage it to entrench dominance. Indian law therefore requires a framework that looks beyond copyright to competition remedies.

The Competition Commission of India (CCI) has gradually recognised the competitive value of data. In the NSE case, proprietary trading access was treated as an essential facility, while in its order against Meta the CCI penalised tying of WhatsApp and Facebook user data. These developments show that Indian regulators can treat data access as a competition issue. However, the doctrine remains underdeveloped, and clear criteria are lacking to decide when denial of data amounts to abuse. Existing statutes also lag: the IT Act and the Digital Personal Data Protection Act exclude public data from restrictions, meaning most scraping is legally permitted. The draft Digital Competition Act proposes rules for "systemically significant digital enterprises," restricting their use of third-party business data and mandating fair conduct, but it does not yet provide specific provisions on scraping.

[45] *Eastern Book Company v D.B. Modak* (2008) 1 SCC 1.
[46] *Ani Media (P) Ltd. v Open AI Inc* 2024 SCC OnLine Del 8120.

ISSN: 2582-6433

Artificial intelligence magnifies these concerns. AI models require vast and diverse datasets, and scraping has become the main source of training material. Once data is ingested, its influence on models becomes permanent, raising issues of privacy, consent, and competitive advantage. Without intervention, incumbents with the ability to scrape and process data at scale could entrench their market dominance, locking out smaller rivals. The current legal position, where public data can be freely scraped, encourages innovation but risks leaving unchecked concentration in the hands of a few platforms.

Comparative experience provides useful lessons. The European Union's Digital Markets Act,[47] imposes proactive obligations on gatekeepers, forbidding them from combining data across services without explicit consent and mandating certain forms of data sharing. EU courts have also been willing to restrict scraping where it undermines competition or database rights. The United States, in contrast, has largely permitted scraping of public data, with *hiQ Labs v. LinkedIn,*[48] confirming that public profiles are not "unauthorised access." US regulators rely mainly on merger review and antitrust enforcement to address data-driven harms. These contrasting models suggest that India must balance openness for innovation with targeted restrictions on dominant firms.

India's approach is characterised by flexibility but also uncertainty. The liberal stance on public data promotes AI research and start-up growth, yet it lacks mechanisms to prevent data hoarding or unfair leveraging by incumbents. Regulatory responsibility is also fragmented, spread across the CCI, the Data Protection Board, and sectoral regulators. Without coordination, there is a risk of inconsistent rulings and gaps in enforcement.

The project therefore suggests several reforms. First, India should clarify the line between permissible and impermissible scraping. Laws could expressly decriminalise collection of open data while prohibiting scraping that circumvents technical barriers or violates privacy safeguards. Second, the Digital Competition Act should be strengthened to impose obligations on dominant platforms: requiring separate, informed consent for combining data and prohibiting the use of non-public business-user data for self-preferencing. Third, the CCI should develop a doctrine of "data essential facilities," identifying circumstances where denial of access to unique datasets constitutes abuse of dominance. Merger review should also

[47] European Union Digital Markets Act, 2022 (2022/1925).
[48] *HiQ Labs, Inc., v Linkedin Corporation*, No. 17-16783 (2022).

ISSN: 2582-6433

explicitly assess data concentration and allow for remedies such as interoperability or limited data-sharing mandates.

Fourth, India needs stronger regulatory coordination. A joint digital markets council could align the CCI, the Data Protection Board, and sectoral regulators, ensuring consistency and avoiding overlap. Fifth, interoperability and portability should be promoted through open APIs and a consumer data right framework, enabling users to move data across platforms. This would reduce lock-in while keeping markets contestable. Finally, issuing best practice guidelines on responsible scraping, for example, respecting robots.txt protocols and applying rate limits, would create predictability for both firms and regulators.

Therefore, India must adopt a balanced approach that encourages beneficial scraping of public data while preventing its misuse by dominant firms. By blending the EU's rule-based safeguards with the US's outcome-focused flexibility, India can create a framework that promotes innovation without tolerating monopolistic data entrenchment. Ultimately, effective reform will require a combination of legislative change, regulatory coordination, and technical guidance to ensure that digital markets remain fair, open, and competitive in the age of artificial intelligence.