International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

The Friction between Generative AI Training and Data Minimization under the DPDP Act, 2023

Mr. Uday Chauhan

ABSTRACT

India has become a hub of technological innovation due to the exponential growth of Generative Artificial Intelligence (AI). But the operational architecture of Large Language Models (LLMs) which require ingestion of massive unfiltered datasets is fundamentally at odds with the data protection principles enshrined under the Digital Personal Data Protection (DPDP) Act, 2023. In this paper we examine the systemic tension between the data consumption requirements of generative machine learning models and the statutory requirement of "Data Minimization" under Indian law. This paper analyses the mechanics of AI ingestion vis-a-vis Section 6 of the DPDP Act and, through a comparative jurisprudence under the GDPR, explores the pitfalls of public data exemptions and demonstrates that the static regulatory framework creates an unsustainable compliance deadlock. This paper finally proposes a reconciled regulatory model with dynamic safe harbors, synthetic data standards, and AI-specific guidelines to harmonize technological advancements with fundamental privacy rights.

INTRODUCTION AND RESEARCH PROBLEM

Today the world economy is undergoing a paradigm shift of unprecedented proportions, powered by the meteoric rise of Generative Artificial Intelligence (AI). The "AI boom", driven by simple Large Language Models (LLMs) and diffusion architectures, has evolved from a conceptual milestone in computer science to a fundamental pillar of industrial and economic infrastructure. India's state-sponsored digital public infrastructure projects and a vibrant startup ecosystem that seeks to make AI-driven solutions accessible to all in health, finance and governance are strongly encouraging this technology boom in India. The structure of Generative AI, however, presents a structural challenge to the traditional legal framework. Unlike traditional software that is explicitly programmed, generative AI models learn through pattern recognition based on data ingestion. The construction of these models depends entirely on the extraction, scraping and processing of huge, unfiltered datasets - billions of data points often picked indiscriminately from the public and private digital spheres. This engineering need is challenged by emerging global and domestic data sovereignty frameworks.

The Research Problem

This paper addresses the acute systemic friction and legal uncertainty arising from the application of the static, transaction-oriented provisions of the Digital Personal Data Protection (DPDP) Act, 2023 to the dynamic, non-linear processing models of machine learning.

First, the basic statutory framework is not regulatory agile. The DPDP Act relies excessively on a rigid, consent-centric client-server processing model that assumes that data collection is always intentional, linear, and predictable.

Second, the statutory requirement of data minimization creates a fundamental, mathematical barrier to modern deep learning. Under the standard data protection doctrine, a data fiduciary cannot collect or hold more data than is strictly necessary for a stated, pre-defined purpose. The future utility, semantic contextualization, and emergent behaviors of a neural network cannot be neatly isolated at the point of web scraping, meaning that AI developers will not be able to comply with static collection limits. This structural incompatibility puts local algorithmic innovation at odds with state privacy mandates, and threatens to disrupt both technological expansion and individual fundamental rights.

RESEARCH OBJECTIVES

The specific, measurable objectives of this doctrinal research project are:

1. To identify the systemic, technical and operational failures that arise when the static data processing definitions of the DPDP Act, 2023 are applied to generative neural networks.
2. To conduct a comparative structural and statutory study of the data minimization and consent requirements under Section 6 of the DPDP Act vis a vis the "Legitimate Interest" routes under Article 6 of the European Union's General Data Protection Regulation (GDPR).
3. Evaluate the legal soundness and structural vulnerabilities of the "Publicly Available Data" exemption under Section 3(c)(ii) of the DPDP Act, with respect to mass machine learning scraping.
4. To study the judicial doctrines laid down by the Supreme Court of India relating to "complete codes", "special economic legislations" and "contextual integration" to determine the interpretive limits of the Data Protection Board (DPB) of India.
5. To identify potential statutory loopholes and formulate policy recommendations for Indian AI developers to establish flexible regulatory safe harbors without sacrificing data privacy.

RESEARCH QUESTIONS

This paper aims to answer the following research questions:

1. To what extent do the strict norms of 'Purpose Limitation' and 'Data Minimisation' under the DPDP Act, 2023 hamper local machine learning engineering and result in compliance deadlocks?
2. What is the structural disadvantage faced by Indian software innovators because of the omission of a private commercial "Legitimate Interest" clause in Section 7 of the DPDP Act vis-à-vis their global counterparts?
3. What are the legal risks, violations of contextual integrity, and liabilities to third parties of webscraping tools harvesting public data under the assumed immunity of Section 3(c)(ii)?
4. What has been the traditional approach of the Supreme Court of India in construing self-contained economic legislations (as in Innoventive Industries, Swiss Ribbons and Girnar Traders) and how do these principles govern the regulatory jurisdiction of the Data Protection Board?

RESEARCH HYPOTHESES

The following legal hypotheses guide the study:

1. **Hypothesis I**: Strict adherence to the data minimization principle under the DPDP Act, 2023 would result in an absolute structural deadlock with respect to Generative AI development, rendering the lawful training of domestic foundation models commercially unviable.
2. **Hypothesis II**: Developers are compelled to rely heavily on volatile exemptions for public data, since the private commercial "legitimate uses" are statutorily excluded under Section 7, exposing themselves

to significant systemic legal liabilities for third-party data contamination.

3. **Hypothesis III**: The Supreme Court's jurisprudence on modern economic regulation authorizes the state, under its secondary rule-making powers, to adopt adaptive technical rules (e.g., differential privacy and machine-readable opt-outs) to balance economic utility and core privacy interests.

RESEARCH METHODOLOGY

In accordance with the guidelines set forth, this project utilizes a doctrinal and qualitative research design.

Data Sources and Material Collection

The qualitative legal data used in the research is obtained from popular sources like Manupatra, SCC Online, HeinOnline and Westlaw India. The evaluated source materials are as follows.

1. Enactment- Reports of previous Committees and the Digital Personal Data Protection Act, 2023 (Act No. 22 of 2023).
2. Comparative frameworks: the General Data Protection Regulation (GDPR) and the Regulation (EU) 2016/679 of European Union.
3. Judicial precedents: Landmark pronouncements of the Supreme Court of India on the rights to privacy, economic experiment, comprehensive code and statutory interpretation.
4. Secondary Literature: Treatises on computer science and the design of neural networks, administrative circulars of international bodies on data protection, and peer-reviewed academic articles.

Analytical Framework

The qualitative data is analyzed using textual, contextual and comparative legal analysis. Structural deviations are identified by way of direct mapping of the engineering processes of deep learning to the thresholds of compliance of the DPDP Act. Simultaneously, the relevant Supreme Court judgments are analysed with the help of case analysis frameworks (the ratio decidendi and obiter dicta) to understand the scope of administrative flexibility available for the Data Protection Board.

LITERATURE REVIEW

The juristic literature, when critically examined, displays a growing global consensus that traditional data protection models are structurally inadequate for the regulation of artificial intelligence systems. The fundamental tension is the transition from transaction processing to algorithmic digestion. Helen Nissenbaum (2010) has suggested the doctrine of "Contextual Integrity" where personal data is inherently connected to the context in which it was originally shared. This is why modern legal scholars argue that the use of public data to power a wholly different commercial AI profit engine violates basic expectations of privacy, no matter what the public availability.

The hallmark of legitimate legal writing is that it does not simply parrot secondary commentary, but instead analyzes the primary legislation directly. As per legal commentators on the Indian tech scene post 2023, the DPDP Act was essentially drafted from a static client-server perspective. Within the Ministry of Electronics and Information Technology (MeitY), proponents say the consent mechanism's absolute clarity enhances user sovereignty.

In contrast, recent critical commentary in technology law blogs (e.g., The Leap Blog, 2026) warns that the Act, by not containing a flexible commercial "legitimate interest" valve, creates a too-centralized, rigid compliance regime. Tech commentators note that European authorities have actively penalized companies

for unmitigated algorithmic scraping, signaling an imminent wave of litigation in India if the statutory text remains unadjusted.

RESEARCH & ANALYSIS

The transition from legacy electronic data storage to decentralized generative networks marks a major technological leap. This section conducts a detailed doctrinal analysis of the specific legal touchpoints where the DPDP Act, 2023 collides with machine learning mechanics.

1. Comparative structural analysis and definitions extension

The DPDP Act divides data processing between the "Data Principal" (the person) and the "Data Fiduciary" (the entity which determines the purpose of processing). 6 of the Act stipulates that consent must be free, specific, informed, unconditional and unambiguous. It requires a specific notice to accompany the request for consent, stating the exact categories of data collected and the specific purpose for which it will be processed.

The Computational Contradiction

Machine learning employs deep learning architectures, where an algorithm scans through huge corpora of text or images, changing millions of internal mathematical weights. This process requires scaling. An AI developer is then limited in the training data it can use, or it sets strict boundaries around narrow, pre-curated datasets, which leads to "underfitting" – where the model is unable to understand contextual nuance, generate coherent output, or perform generalized tasks.

This structure exposes three distinct legal dilemmas:

1. The Specificity Problem. An AI creator scraping open web links cannot realistically give prior, specific notice to millions of internet users before their data points are ingested into a training pipeline. This runs counter to the sequential requirement of notice before acquisition.
2. Purpose Indeterminacy: Feeding data to an LLM does not do a single linear task. The eventual deployment of a trained model is infinitely variable, and cannot be defined or limited at the time of the initial data scraping.
3. The Erasure Paradox: Once personal data has been learned into the structural weights of a trained model, it is mathematically impossible to selectively "delete" a single individual's data without retraining the full model from scratch at enormous economic cost. This runs counter to the right to erasure under Section 12 of the DPDP Act.

2. The Omission of "Legitimate Interest" and Global Disparities

The conflict between data minimization and AI is an international battleground. Looking at the European Union's General Data Protection Regulation (GDPR) provides vital guidance on this friction. Article 5(1)(c) GDPR expressly provides for the principle of data minimization. European data protection authorities have been notoriously strict on this rule, for example the Italian Data Protection Authority (Garante per la protezione dei dati personali) temporarily suspended ChatGPT in 2023 on the grounds that there was no legal basis for mass algorithmic data collection.

However, the GDPR has a regulatory safety valve, which is absent in the Indian DPDP Act: the doctrine of "Legitimate Interest" under Article 6(1)(f). This allows European AI companies to argue that they can process public data for AI training without explicit consent, provided that the company's legitimate commercial or research interests are not overridden by the fundamental rights of the data subjects.

In sharp contrast, the Indian DPDP Act, 2023 deliberately excluded a broad "legitimate interest" clause for private entities. Instead, the Act introduces the idea of "Certain Legitimate Uses" in Section 7. A

careful reading of Section 7, however, reveals that these legitimate uses are narrowly defined, including voluntary disclosure of data, state security functions, medical emergencies and employment purposes. Importantly, the DPDP Act does not contain any legislative provision that allows a private tech corporation to ingest personal data for the purpose of commercial algorithmic training under the guise of a "legitimate use."

3. The Publicly Available Data Exemption and Its Structural Pitfalls

Confronted with this statutory wall, AI developers often turn to other legal arguments to support mass web scraping. The most common defense is the "Publicly Available Data" argument. The logic being that if the data principal has voluntarily published their personal data on public forums, social media platforms or open blogs, they have effectively waived their expectation of privacy and therefore the data is exempt from DPDP compliance under Section 3(c)(ii).

But there are three core pitfalls to using this exemption for Generative AI training that make it a legal minefield:

The Contextual Integrity Breach A data principal might upload a resume to LinkedIn to network professionally, or share a personal story on a public blog to express themselves creatively. This voluntary act is defined by contextual integrity. The user didn't make that data public so that some commercial enterprise could scrape it, vectorize it, and use it to build a commercial AI profit engine that might displace the user's own livelihood one day. The DPDP Act text does not explicitly provide a license for commercial exploitation of public data.

B. The Taint of the Third-Party Web scraping tools are inherently indiscriminate. When an AI company scrapes a public website, it collects data that was uploaded by the data principal, and not only that. It also mines personal data uploaded by third parties without the principal's consent (e.g. a blog post written by Person A disclosing private details about Person B). In this case 3(c)(ii) fails completely as the data principal did not make the data public. The AI company becomes a holder of the unlawfully processed personal data.

C. State Exemptions and Research Some believe that AI development can hide behind Section 17 of the Act, which gives the Central Government the power to exclude certain data fiduciaries or processing activities from the core provisions of the Act, specifically processing for "research, archiving or statistical purposes". This exemption is, however, subject to strict conditions: the data may not be used to make any decision that is specific to a data principal and the processing must comply with strict standards set by the state. For commercial companies deploying consumer-facing LLMs, the research exception is not a legal loophole for compliance.

4. Judicial Jurisprudence on Special Economic Codes and Regulatory Flexibility

Much of the legal life and operational limits of the DPDP Act hinge on the rich constitutional and administrative jurisprudence of the Supreme Court of India. In judging the validity of fast-track technical regulations and administrative choices, certain judicial canons arise which give the state executive leeway to govern complicated economic and technological environments.

A. The Broad Statutory Boundary Rule The Supreme Court has consistently held that modern, consolidated economic legislations should be interpreted on their own contemporary text, unburdened by legacy legal constraints. In Innoventive Industries Ltd. v. ICICI Bank and Another, the Court clarified that consolidation and amendment of financial enactments is a complete code in itself. The Court observed

that the main aim of such all-embracing codes is to prevent defaulting parties from using external, general legislations to slow down or obstruct the enforcement of specialized statutory mandates.

B. The Doctrine of Complete Machinery and Administrative Discretion

The structural insulation was further strengthened by the Constitution Bench in Girnar Traders v. State of Maharashtra and Others. The Court noted that a self-contained code contains all the necessary internal machinery to achieve its designated social or economic purpose. To read external, unaligned procedures or limitations into a highly specialized statutory scheme would be to defeat the very legislative intent of creating a unified regulatory perimeter.

C. Judicial Deference to Economic and Technological Experimentation

Most importantly, in Swiss Ribbons Pvt. Ltd. and Another v. Union of India and Others, the Supreme Court dismissed constitutional challenges to specialized regulatory regimes, stating that the legislature and statutory boards should be given enough room to experiment with laws on economics to ensure stability in the market and national development. Jurisdiction in this jurisprudence directly supports the rule-making powers of the Data Protection Board of India.

The Supreme Court recognizes that specialized administrative panels have the legal capacity to prepare dynamic rules, exemptions, and technological standards to implement the main text of a master code. Therefore, the Data Protection Board can lawfully exercise its secondary law-making powers to prescribe AI-specific technical standards without breaching the basic separation of powers.

SUGGESTIONS AND RECOMMENDATIONS

To resolve the statutory gaps, structural ambiguities, and compliance deadlocks revealed in this doctrinal analysis, the following legislative and administrative proposals are advanced:

1. Rule Making Powers to Create Dedicated AI Training Safe Harbors: The Data Protection Board of India should create specific rules to create a “AI Training Safe Harbor.” This mechanism would enable the use of publicly available personal data for foundation model training without explicit consent, provided the AI developer applies strict automated filtering to eliminate highly sensitive personal data (e.g., financial data, health data, children’s data)prior to the training phase.
2. Statutory Standards for Differential Privacy and Synthetic Data: Regulators should focus on reducing the identifiability of the data, not the amount of data collected, which harms AI utility. The government should codify advanced cryptographic and statistical techniques (e.g. Differential Privacy) as acceptable compliance measures. In addition, the state should incentivize the creation and deployment of fully synthetic datasets: datasets that are artificially generated by algorithms that mimic the statistical properties of real-world data, but without containing any actual personal information of real data principals.
3. Technical Frameworks for Algorithmic “Opt-Outs”: India must develop an enforceable technical standard for algorithmic opt-outs that respects individual autonomy without requiring economically disastrous model-retraining cycles. Like the "Robots.txt" protocol used by the webmasters to block search engine crawlers, a standard metadata tag should be legally recognized under the DPDP rules. That would allow data principals to signal in a machine-readable format that their public digital footprints should not be harvested for machine-learning purposes. Any AI developers that breach this express opt-out would face the maximum statutory penalties.

CONCLUSION

The Digital Personal Data Protection Act, 2023 is a milestone for Privacy Rights in India. But its architecture is still rooted in the age of linear data transactions. Generative AI upsets this foundation with the realization that progress in technology depends on aggregating data while modern privacy law requires an abundance of data minimization.

If India wants to be a world leader in artificial intelligence, the continued existence of a rigid, inflexible interpretation of Section 6 and data minimization principles will lead to an unsustainable legal environment. It puts AI developers in a legal grey area, and it gives citizens illusory protections. India's Data Protection Board can successfully bridge this divide by embracing dynamic safe harbors, technological privacy standards like differential privacy and strong opt-out frameworks. The answer is not to sacrifice privacy or to stifle innovation but to establish a nimble, tech-savvy regulatory framework within which both can flourish at the same time.

REFERENCES

Primary Sources

Statutes:

1. Digital Personal Data Protection Act, 2023, No. 22, Acts of Parliament, 2023 (India).
2. Council Regulation 2016/679, General Data Protection Regulation (GDPR), 2016 O.J. (L 119) 1 (EU).

Judgments:

1. Girnar Traders v. State of Maharashtra and Others, (2011) 3 SCC 1.
2. Innoventive Industries Ltd. v. ICICI Bank and Another, (2018) 1 SCC 407.
3. Swiss Ribbons Pvt. Ltd. and Another v. Union of India and Others, (2019) 4 SCC 17.
4. Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.

Secondary Sources

Reports & Treatises:

1. Helen Nissenbaum, Privacy in Context: Technology, Policy, and the Integrity of Social Life (Stanford University Press, 2010).
2. Ian Goodfellow, Yoshua Bengio, & Aaron Courville, Deep Learning (MIT Press, 2016).
3. Cynthia Dwork & Aaron Roth, The Algorithmic Foundations of Differential Privacy, 9(3–4) Foundations and Trends in Theoretical Computer Science 211 (2014).

Articles & Commentaries:

1. Kapp, Marshall B., Writing Research Papers: 10 Top Tips, 17(3) The Law Teacher (1999).
2. Garante per la protezione dei dati personali, Provision of 30 March 2023 regarding ChatGPT algorithmic scraping, Provision No. [9870832], Italian Data Protection Authority (2023).
3. Leap Blog Contributors, Comments on the Digital Personal Data Protection Compliance Architecture, The Leap Journal (April 2026).