THE FRICTION BETWEEN GENERATIVE AI TRAINING AND DATA MINIMIZATION UNDER THE DPDP ACT, 2023

Vedanti Rajput, Bharatividyapeeth Institute of Management and Research

ABSTRACT

The exponential growth of Generative Artificial Intelligence (AI) has positioned India as a central hub for technological innovation. However, the operational architecture of Large Language Models (LLMs)—which require the ingestion of massive, unfiltered datasets—fundamentally conflicts with the data protection principles established under the Digital Personal Data Protection (DPDP) Act, 2023. This paper examines the systemic friction between the data consumption needs of generative machine learning models and the statutory mandate of "Data Minimization" under Indian law. By analyzing the mechanics of AI ingestion against Section 6 of the DPDP Act, exploring comparative jurisprudence under the GDPR, and evaluating the pitfalls of public data exemptions, this paper demonstrates that the current static regulatory framework creates an unsustainable compliance deadlock. Ultimately, this paper proposes a reconciled regulatory model featuring dynamic safe harbors, synthetic data standards, and AI-specific guidelines to balance technological progress with fundamental privacy rights.

INTRODUCTION AND RESEARCH PROBLEM

The contemporary global economy is experiencing an unprecedented paradigm shift driven by the meteoric rise of Generative Artificial Intelligence (AI). Driven by foundational Large Language Models (LLMs) and diffusion frameworks, the "AI boom" has transitioned from a theoretical computer science milestone into a core pillar of industrial and economic infrastructure. In India, this technological surge is heavily incentivized by state-backed digital public infrastructure initiatives and a thriving startup ecosystem aiming to democratize AI-driven solutions across healthcare, finance, and governance.

However, the operational architecture of Generative AI poses a structural challenge to traditional legal frameworks. Unlike classical software, which relies on explicit programming, Generative AI models learn via pattern recognition derived from data ingestion. The development of these models relies entirely on the extraction, scraping, and processing of vast, unfiltered datasets—billions of data points frequently harvested indiscriminately from public and private digital spheres. This engineering necessity directly collides with emerging global and domestic data sovereignty frameworks.

The Research Problem

This research 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 core statutory framework lacks regulatory agility. The DPDP Act relies heavily on a rigid, consent-centric client-server processing paradigm that assumes data collection is always deliberate, linear, and predictable.

Second, the statutory mandate of Data Minimization poses a fundamental, mathematical obstacle to modern deep learning. Under standard data protection doctrine, a data fiduciary is forbidden from collecting or retaining more data than is strictly necessary to achieve a stated, pre-defined purpose. Because the future utility, semantic contextualization, and emergent behaviors of a neural network cannot be neatly isolated at the point of web scraping, AI developers find it operationally impossible to comply with static collection limitations. This structural incompatibility places local algorithmic innovation in direct conflict with state

privacy mandates, threatening to disrupt both technological expansion and individual fundamental rights.

RESEARCH OBJECTIVES

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

1. To identify the systemic, technical, and operational failures that arise when applying the static data processing definitions of the DPDP Act, 2023 to generative neural networks.

2. To carry out a comparative structural and statutory analysis of the data minimization and consent mandates under Section 6 of the DPDP Act vis-à-vis the "Legitimate Interest" pathways under Article 6 of the European Union's General Data Protection Regulation (GDPR).

3. To evaluate critically the legal validity and structural pitfalls of relying on the "Publicly Available Data" exemption under Section 3(c)(ii) of the DPDP Act for mass machine learning scraping.

4. To study the judicial doctrines laid down by the Supreme Court of India in respect of "complete codes," "special economic legislations," and "contextual integration" to evaluate the interpretive boundaries of the Data Protection Board (DPB) of India.

5. To identify potential statutory loopholes and formulate actionable policy suggestions to establish flexible regulatory safe harbors for Indian AI developers without compromising data privacy.

RESEARCH QUESTIONS

This paper addresses the following research questions:

1. To what extent does the rigid framework of "Purpose Limitation" and "Data Minimization" under the DPDP Act, 2023 impede local machine learning engineering and promote compliance deadlocks?

2. How does the deliberate omission of a private commercial "Legitimate Interest" clause in Section 7 of the DPDP Act structurally disadvantage Indian software innovators compared

to their global counterparts?

3. What legal risks, contextual integrity violations, and third-party liabilities arise when web-scraping tools harvest 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 study is guided by the following legal hypotheses:

1. **Hypothesis I:** The literal application of the data minimization principle under the DPDP Act, 2023 to Generative AI development will result in an absolute structural deadlock, rendering the lawful training of domestic foundation models commercially unviable.

2. **Hypothesis II:** The statutory exclusion of private commercial "legitimate uses" under Section 7 forces developers into an over-reliance on unstable public data exemptions, exposing them to massive systemic legal liabilities for third-party data contamination.

3. **Hypothesis III:** Under the Supreme Court's established jurisprudence on modern economic regulations, the state can introduce adaptive technical rules (such as differential privacy and machine-readable opt-outs) under its secondary rule-making powers to balance economic utility with fundamental 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 study is based on qualitative legal data collected from standard legal databases like Manupatra, SCC Online, HeinOnline, and Westlaw India. The following source materials are

examined:

1. **Primary Legislation:** The Digital Personal Data Protection Act, 2023 (Act No. 22 of 2023) along with historical committee reports.

2. **Comparative Frameworks:** The European Union General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679).

3. **Judicial Precedents:** Landmark judgments of the Supreme Court of India concerning statutory construction, complete codes, economic experimentation, and the fundamental right to privacy.

4. **Secondary Literature:** Academic articles from peer-reviewed journals, administrative circulars from international data protection boards, and computer science treatises on neural network architecture.

Analytical Framework

The qualitative data is analyzed through textual, contextual, and comparative legal analysis. Structural deviations are determined by mapping the engineering processes of deep learning directly against the compliance thresholds of the DPDP Act. Concurrently, relevant Supreme Court judgments are examined using case-analysis frameworks (evaluating the ratio decidendi and obiter dicta) to determine the scope of administrative flexibility available to the Data Protection Board.

LITERATURE REVIEW

A critical examination of juristic literature highlights a growing global realization that traditional data protection models are structurally ill-equipped to govern artificial intelligence systems. The foundational tension lies in the shift from transactional processing to algorithmic ingestion. Helen Nissenbaum (2010) established the doctrine of "Contextual Integrity," arguing that personal data is inextricably tied to the context in which it was originally shared. This underscores why modern legal writers argue that using public data for an entirely separate commercial AI profit engine violates fundamental privacy expectations, regardless of public accessibility.

Legitimate legal writing requires the direct analysis of primary legislation rather than merely paraphrasing secondary commentary. Legal commentators analyzing the Indian tech landscape post-2023 note that the DPDP Act was fundamentally drafted through a static client-server lens. Supporters within the Ministry of Electronics and Information Technology (MeitY) assert that the absolute clarity of the consent mechanism enhances user sovereignty.

Conversely, recent critical commentary in technology law blogs (e.g., The Leap Blog, 2026) cautions that by omitting a flexible commercial "legitimate interest" valve, the Act creates an over-centralized, rigid compliance regime. Tech commentators note that European authorities have actively penalized companies for unmitigated algorithmic scraping, signaling an impending 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 Expansion

Under the DPDP Act, data handling is divided between the "Data Principal" (the individual) and the "Data Fiduciary" (the entity determining the purpose of processing). Section 6 of the Act mandates that consent must be free, specific, informed, unconditional, and unambiguous. It requires a clear notice accompanying the request for consent, specifying the exact data categories collected and the precise purpose of processing.

The Computational Contradiction

Machine learning operates on deep learning architectures, where an algorithm adjusts millions of internal mathematical weights by analyzing massive corpora of text or imagery. This process requires scale. If an AI developer minimizes the training data or restricts it strictly to narrow, pre-curated datasets, the model suffers from "underfitting"—rendering it incapable of understanding contextual nuances, generating coherent output, or performing generalized tasks.

STATUTORY PARAMETER	**TRADITIONAL LINE PROCESSING**	**GENERATIVE AI TRAINING PIPELINE**
Data Ingestion volume	Minimized, limited strictly to specific transaction fields	Maximized, relying on millions of data points for pattern recognition.
Notice Requirement	Prior, explicit notice delivered directly to the individual	Absent, web scraping tools, gather public Web links indiscriminately
Purpose Definition	Unambiguous, static, and predefined.	Indeterminate, data is consumed to teach language structures for infinite future uses.

This structure exposes three distinct legal dilemmas:

1. **The Specificity Dilemma:** An AI developer scraping open web links cannot realistically provide prior, specific notice to millions of internet users before their data points are ingested into a training pipeline. This breaks the sequential requirement of notice before acquisition.

2. **The Purpose Indeterminacy:** Ingesting data into an LLM does not accomplish a single, linear task. The eventual deployments of a trained model are infinitely variable and cannot be defined or limited at the moment of initial data scraping.

3. **The Erasure Paradox:** Once personal data is integrated into the structural weights of a trained model, it becomes mathematically impossible to selectively "delete" a single individual's data without completely retraining the model from scratch at immense economic cost. This directly clashes with the right to erasure under Section 12 of the DPDP Act.

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

The tension 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) of the GDPR explicitly outlines the data minimization principle.

European data protection authorities have taken a notoriously stringent approach to this rule; for instance, the Italian Data Protection Authority (Garante per la protezione dei dati personali) temporarily banned ChatGPT in 2023, citing a lack of legal basis for mass algorithmic data collection.

However, the GDPR possesses a regulatory safety valve that the Indian DPDP Act lacks: the doctrine of "Legitimate Interest" under Article 6(1)(f). This allows European AI companies to argue that processing public data for AI training is permissible 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 stark contrast, the Indian DPDP Act, 2023 deliberately omitted a broad "legitimate interest" clause for private entities. Instead, Section 7 of the Act introduces the concept of "Certain Legitimate Uses." A close reading of Section 7 reveals that these legitimate uses are strictly narrowly tailored, covering scenarios such as voluntary data disclosure, state security functions, medical emergencies, and employment purposes. Crucially, there is no legislative provision within the DPDP Act that permits a private tech corporation to ingest personal data for commercial algorithmic training under the umbrella of a "legitimate use."

3. The Publicly Available Data Exemption and Its Structural Pitfalls

Faced with this statutory wall, AI developers frequently rely on alternative legal arguments to justify mass web scraping. The most common defense is the "Publicly Available Data" argument. The logic follows that if a data principal voluntarily publishes their personal data on public forums, social media platforms, or open blogs, they have effectively waived their expectation of privacy, thereby exempting that data from DPDP compliance under Section 3(c)(ii).

However, relying on this exemption for Generative AI training is a legal minefield due to three core pitfalls:

A. The Contextual Integrity Violation

A data principal may post a resume on LinkedIn for professional networking, or share a personal story on a public blog for creative expression. This voluntary action is bound by contextual integrity. The user did not make that data public so that a commercial enterprise

could scrape it, vectorize it, and use it to train a commercial AI profit engine that might eventually displace the user's own livelihood. The DPDP Act's text does not explicitly grant a blanket license for commercial exploitation of public data.

B. The Third-Party Taint

Web scraping tools are inherently indiscriminate. When an AI company scrapes a public website, it does not only collect data uploaded by the data principal. It also collects personal data uploaded by third parties without the principal's consent (e.g., a blog post written by Person A exposing private details about Person B). In this scenario, Section 3(c)(ii) fails entirely, because the data was not made public by the data principal. The AI company becomes a possessor of unlawfully processed personal data.

C. The State and Research Exemptions

Some argue that AI development can shelter under Section 17 of the Act, which allows the Central Government to exempt certain data fiduciaries or processing activities from the core provisions of the Act—specifically processing for "research, archiving, or statistical purposes." However, this exemption is heavily conditional: the data cannot be used to make any decision specific to a data principal, and the processing must adhere to strict standards prescribed by the state. For commercial entities deploying consumer-facing LLMs, escaping compliance via the research exemption is legally unviable.

4. Judicial Jurisprudence on Special Economic Codes and Regulatory Flexibility

The legal survival and operational boundaries of the DPDP Act rely heavily on the rich constitutional and administrative jurisprudence of the Supreme Court of India. When evaluating the validity of fast-track technical regulations and administrative choices, specific judicial canons come into play, granting the state executive latitude to govern complex economic and technological environments.

A. The Rule of Comprehensive Statutory Perimeters

The Supreme Court has consistently recognized that modern, consolidated economic legislations must be interpreted on their own contemporary text, unburdened by legacy legal constraints. In Innoventive Industries Ltd. v. ICICI Bank and Another, the Court affirmed that

consolidating and amending financial enactments constitute an exhaustive code unto themselves. The Court noted that the central objective of such comprehensive codes is to prevent defaulting parties from utilizing external, general legislations to slow down or impede the enforcement of specialized statutory mandates.

B. The Doctrine of Complete Machinery and Administrative Latitude

This structural insulation was further reinforced by the Constitution Bench in Girnar Traders v. State of Maharashtra and Others. The Court observed that a self-contained code provides complete internal machinery for the execution of its designated social or economic purpose. To read external, unaligned procedures or limitations into a highly specialized statutory scheme would defeat the very legislative intent of creating a unified regulatory perimeter.

C. Judicial Deference to Economic and Technological Experimentation

Crucially, in Swiss Ribbons Pvt. Ltd. and Another v. Union of India and Others, the Supreme Court repelled constitutional challenges to specialized regulatory frameworks, holding that the legislature and statutory boards must be granted adequate latitude to experiment with economic laws aimed at ensuring market stability and national development. This jurisprudence provides direct support to the rule-making powers of the Data Protection Board of India.

The Supreme Court acknowledges that specialized administrative boards possess the legal authority to draft dynamic rules, exemptions, and technological standards to implement the primary text of a master code. Consequently, the Data Protection Board can legally deploy its secondary rule-making powers to introduce AI-specific technical standards without violating the fundamental 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. **Formulate Dedicated AI Training Safe Harbors under Rule-Making Powers:** The Data Protection Board of India should issue targeted regulations establishing an "AI Training Safe Harbor." This mechanism would permit the processing of publicly accessible personal data for foundation model training without explicit consent, provided that the AI developer

implements rigorous automated filtering to strip out highly sensitive personal data (such as financial information, health records, and children's data) prior to the training phase.

2. Statutory Standards for Differential Privacy and Synthetic Data: Rather than attempting to minimize the volume of data collected (which damages AI utility), regulatory focus should shift toward minimizing the identifiability of the data. The government should legally recognize advanced cryptographic and statistical techniques, such as Differential Privacy, as valid compliance measures. Furthermore, the state should incentivize the creation and deployment of fully synthetic datasets—data generated artificially by algorithms that mimic the statistical properties of real-world data without containing any actual personal information of real data principals.

3. Technical Frameworks for Algorithmic "Opt-Outs": To honor individual autonomy without forcing economically ruinous model-retraining cycles, India should implement an enforceable technical standard for algorithmic opt-outs. Similar to the "Robots.txt" protocol used by webmasters to block search engine crawlers, a standardized metadata tag should be legally recognized under the DPDP rules. This would allow data principals to signal, in a machine-readable format, that their public digital footprints must not be harvested for machine learning purposes. AI developers who violate this explicit opt-out would face maximum statutory penalties.

CONCLUSION

The Digital Personal Data Protection Act, 2023 represents a milestone for privacy rights in India. However, its architecture remains anchored in an era of linear data transactions. The advent of Generative AI disrupts this foundation by demonstrating that technological progress relies on data aggregation, while modern privacy law demands data minimization.

As India seeks to anchor its position as a global leader in artificial intelligence, maintaining a rigid, unyielding interpretation of Section 6 and data minimization principles will create an unsustainable legal environment. It forces AI developers to operate in a legal gray zone and leaves citizens with illusory protections. By embracing dynamic safe harbors, technological privacy standards like differential privacy, and robust opt-out frameworks, the Data Protection Board of India can successfully bridge this divide. The path forward lies not in weakening privacy or slowing down innovation, but in building an agile, technologically sophisticated regulatory environment where both can concurrently thrive.

REFERENCES

Primary Sources

Statutes:

• Digital Personal Data Protection Act, 2023, No. 22, Acts of Parliament, 2023 (India).

• Council Regulation 2016/679, General Data Protection Regulation (GDPR), 2016 O.J. (L 119) 1 (EU).

Judgments:

• Girnar Traders v. State of Maharashtra and Others, (2011) 3 SCC 1.

• Innoventive Industries Ltd. v. ICICI Bank and Another, (2018) 1 SCC 407.

• Swiss Ribbons Pvt. Ltd. and Another v. Union of India and Others, (2019) 4 SCC 17.

• Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.

Secondary Sources

Reports & Treatises:

• Helen Nissenbaum, Privacy in Context: Technology, Policy, and the Integrity of Social Life (Stanford University Press, 2010).

• Ian Goodfellow, Yoshua Bengio, & Aaron Courville, Deep Learning (MIT Press, 2016).

• Cynthia Dwork & Aaron Roth, The Algorithmic Foundations of Differential Privacy, 9(3–4) Foundations and Trends in Theoretical Computer Science 211 (2014).

Articles & Commentaries:

• Kapp, Marshall B., Writing Research Papers: 10 Top Tips, 17(3) The Law Teacher (1999).

• Garante per la protezione dei dati personali, Provision of 30 March 2023 regarding ChatGPT algorithmic scraping, Provision No. [9870832], Italian Data Protection Authority (2023).

• Leap Blog Contributors, Comments on the Digital Personal Data Protection Compliance Architecture, The Leap Journal (April 2026).