JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AI TRAINING: STRENGHTHENING DPDP ACT TO ENHANCE DATA PRIVACY

AUTHORS:
SHREELAKHSHMI K, APARNA ARUN
5th year; B.Com.,LL.B (Hons.), 5th year B.A.,LL.B (Hons.)
School of Law, SASTRA Deemed University, Thanjavur

ABSTRACT

Every powerful Artificial Intelligence model is fueled by enormous amounts of personal data, archives of human experiences often harvested without user consent. This silent collection of data is used to feed machine learning algorithms to enable them to identify patterns and to refine their predictive capabilities which raises concerns of privacy infringement. This paper focuses on AI systems in India, the techniques used by it to acquire user data and the lack of transparency in data accumulation processes. Further, these datasets are used for training AI models to enhance predictive analytics, targeted advertising, and automated decision-making. AI Models actively collect countless interactions and transactions, perpetually generating vast databases for training purposes. It studies whether the Digital Personal Data Protection Act (DPDP), 2023 is sufficient to safeguard users from exploitative AI training methods. This paper compares different international frameworks such as the European Union's General Data Protection Regulation (GDPR) and other Data Privacy Laws which emphasises stronger consent, transparency and accountability measures. This paper seeks to contribute to enhancing Responsible AI in India by offering suggestions. This includes strengthening informed consent mechanisms and strong privacy protection regime and investing in user awareness programmes to ensure AI development which aligns with dignity, trust and autonomy.

KEYWORDS

AI Training, Data Privacy, Privacy Infringement, Publicly Available Personal Data

I. INTRODUCTION TO AI TRAINING

AI has now become an inseparable part of our everyday life from ChatGPT, Perplexity, and Gemini generating text to facial recognition unlocking phones, healthcare AIs diagnosing diseases and applications guiding routines. The accuracy, fairness, and reliability of these systems depend directly on the quality and diversity of the data used to train them. These data enable these systems to interpret patterns, generate content, and make decisions that align closely with human expectations.AI training means the process of teaching an AI to refine its capabilities by repeatedly processing large volumes of data. AI training data consists of structured information and inputs that enable AI models to make reliable predictions and decisions. AI training data comes in many forms depending on the model's purpose. Text data from sources like books, websites, and research papers teaches language models such as ChatGPT, Perplexity, and Gemini to understand and generate human language. Audio data trains voice assistants and speech-to-text systems to recognize accents, tones, and even emotions, while also processing sounds like music. Image and video data support computer vision applications used in facial recognition, medical imaging, surveillance, and autonomous vehicles. AI models can be trained using labeled, unlabeled, or mixed data. Labeled data includes tags or annotations that guide the model, a process known as supervised learning. Unlabeled data lacks such tags and is used in unsupervised learning to let the AI find patterns on its own. Combining both types creates more balanced and effective AI models. The other is reinforcement learning, where an AI model learns by performing actions and receiving feedback in the form of rewards or penalties. This process enables the model to understand the outcomes of its choices and gradually improve its decision-making over time.[i]

II. DATA COLLECTION TECHNIQUES USED BY AI MODELS

The effectiveness of AI models is fundamentally grounded in the input data used during their development, which enables them to learn complex relationships, recognize patterns, and subsequently generate coherent content, make accurate predictions, and

understand context. The Performance of the AI model scales directly with data volume, necessitating vast quantities of data. However, sheer quantity must be balanced by data quality.

The methods of collecting or generating data for machine learning and AI applications include:

- **Web Scraping**

Web scraping collects data for AI training by deploying automated bots or scripts to systematically extract and aggregate vast quantities of text, images, and other content from numerous public websites. This process breaches privacy when the collection is indiscriminate, encompassing personally identifiable information (PII) like names, contact details, private communications, or sensitive attributes that individuals did not explicitly consent to have used for commercial AI development. Although the data may be publicly accessible, its large-scale, unauthorized harvesting for training Large Language Models (LLMs) or other AI systems. The legal and tension is exemplified in cases like ***Reddit, Inc. v. Anthropic***, ***PBC***, where Reddit alleges Anthropic scrapped its platform without permission to train and commercialize its Claude AI model, violating Reddit's User Agreement and leading to claims including breach of contract and unfair competition. Due to increasing opposition, some large data aggregators use a subtle workaround: they contract directly with a platform's end users (individual account holders) for access. This involves asking users to "link their account" by providing credentials, allowing the aggregator to collect proprietary data either by scraping the site using the customer's access or via an authorized API. The aggregator claims its access is lawful because the customer consented, sidestepping cybersecurity laws.[ii]

- **APIs and Public Datasets**

Public Datasets are structured, machine-readable data published by organizations and governments in data repositories and portals, typically under an open license. This information can be collected using Application Programming Interfaces (APIs), which are protocols allowing software systems to communicate and exchange data efficiently. Commercial entities, like Yahoo , Google , also use APIs to grant controlled access to proprietary datasets. Many platforms, like X (formerly Twitter), offer public APIs as a reliable, structured way to access data, though these often impose rate limits or require access fees. Many platforms are adopting a strategic and proactive approach by channelling access through API agreements. This establishes a secure gateway, allowing third parties to access specific data fields under defined conditions. These agreements integrate necessary guardrails for security, usage, and compliance, enabling the host to impose restrictions, track data usage, and effectively mitigate risks of unauthorized access. A crucial consideration when using these public APIs is the inherent lack of direct, individual user consent for secondary data use, as the platform itself is granting access to the content generated by its users.[iii]

- **User-Generated Data**

AI models are increasingly trained on real-world, dynamic data sourced directly from users and connected devices, highlighting issues of user privacy and explicit consent. Companies like Meta publicly state their intention to use public posts, comments, and AI queries to train their models. This approach, exemplified by companies like Zomato and Swiggy in India using customer transaction data, often involves platforms asserting a right to use data under broad Terms of Service agreements rather than seeking fresh, granular consent for AI training, raising concerns about the scope of data usage beyond a user's initial expectations. Devices embedded with sensors collect valuable operational data. Tesla, for instance, uses anonymized real-world driving data collected by in-vehicle sensors to train its self-driving system. Similarly, in India, firms using farm sensors or medical diagnostics collect data where the individual's awareness or control over its secondary use for AI model development is often minimal or entirely absent. This constant, streamable data underscores a critical tension between the demand for highly effective AI and the individual's privacy and informed consent.

III. CONSENT MECHANISM, CONSEQUENCES OF NON-CONSENSUAL DATA REPURPOSING

The consent mechanism for AI should be specific, informed, and unambiguous, giving the user a genuine choice about whether their personal content (e.g., public posts, transaction history) can be used for purposes like model development. When this informed consent is lacking, and data is repurposed without the user's explicit knowledge, the consequences are severe. The non-consensual repurposing of data for AI training, often through broad Terms of Service or unauthorized scraping, creates significant legal consequences. Furthermore, this practice erodes public trust, as users feel deceived when their data is used beyond their initial expectation. Bypassing consent for AI training undermines user autonomy and presents a foundational risk to the responsible development and deployment of AI technology. When using ChatGPT, user conversations are typically used for AI model training by default, with the burden placed on users to opt out, rather than requiring proactive, explicit consent before data collection begins. Most users remain unaware that their chats often containing sensitive information are retained and processed for ongoing AI development highlighting persistent risks of silent data harvesting and inadequate transparency in the current regulatory environment. [iv]A Digital Digging investigation analysed 512 publicly shared ChatGPT conversations using targeted keyword searches. The rapid growth of generative AI in India has introduced both innovation and serious privacy risks, as highlighted by the investigation by Digitaldigging.org [v]into the "ChatGPT Confession Files." The probe revealed that AI platforms' "share" features inadvertently made confidential business, legal, and personal conversations publicly searchable, exposing sensitive data such as proprietary code, corporate strategies, and personal health information.Many professionals believed their shared chats were private, but a "discoverable" toggle allowed these conversations to be indexed by search engines, sometimes even after deletion. OpenAI has since removed this feature, but the incident underscores how quickly design flaws or user misunderstandings can lead to severe data exposure, especially in India, where global clients rely on strict privacy standards.

IV. INDIAN LEGAL APPROACH ON DATA PRIVACY

The Digital Personal Data Protection Act, 2023 (DPDP Act)[vi] creates India's comprehensive framework for processing of digital personal and data protection. It applies to data processed within India and to data processed outside India when related to services offered to Indian residents. The DPDP Act's general approach to data protection creates critical gaps that AI companies can leverage for data collection, particularly through interpretative ambiguities and the lack of explicit AI-focused approach. AI training falls under the Act's processing of personal data, and thus is governed by a consent-centric regime. The Section 4 of the DPDP Act recognises "consent and certain legitimate uses" as two grounds for processing personal data and Section 7 specifies the scenarios where personal data can be processed without the consent of the Data Principal. These include situations where Data Principal has voluntarily provided their personal data to the Data Fiduciary and has not objected to its use for a specific purpose, medical emergencies, employment purposes, safety measures during disaster or breakdown of public order, compliance with any judgement or decree, for fulfilling any obligation under the law and provisions of any government services and benefits.

In case of Section 6 the consent shall be free, specific, informed, unconditional and unambiguous and shall signify an agreement to the processing of personal data for the specified purpose.

The DPDP Act has the following shortcomings in including the regulation of AI Training: -

- Section 17(2)(b) of the act exempts the processing of data from the scope of this act if it is necessary for research, archiving or statistical purposes. AI companies can exploit this loophole by claiming their for-profit model development as "research." This allows them to collect huge amounts of data without needing full user consent.
- Section 6 also deals with collection of data for specific purpose which is vulnerable to "function creep." A company may initially collect data for a narrow purpose, such as service improvement. Over time, it can then repurpose this accumulated large dataset to train a new commercial generative AI agent like targeted advertising, justifying the move by claiming the AI aligns with the original, broadly stated goal of "service improvement." Without clear limits on data repurposing for AI, this gap allows companies to exploit existing data reserves for new, unanticipated commercial applications.[vii]
- A significant regulatory gap for AI training relates to the use of data that is already public. Section 3 of the act exempts the data that is made or caused to be made publicly available from the scope of this act. AI companies frequently engage in collection of massive quantities of personal data from public platforms, social media, and scrapped websites. They operate under the assumption that since the data was voluntarily disclosed by the user to the public, its collection and use for commercial AI training often framed as a public benefit or scientific advance falls under a "legitimate use" exemption. The difficulty here is that the context of public disclosure differs vastly from the purpose of commercial scraping. For example, A thought posted on a blog being monetized for developing a proprietary model. Crucially, the Act lacks specific provisions that govern whether making data public inherently grants an unlimited license for commercial, for-profit training of AI models, thus leaving a major loophole that enables the mass extraction and misuse of user information.

In summary, the DPDP Act establishes core protections for personal data and introduces exemptions that could, in principle, accommodate certain AI training activities, particularly when data is publicly available or when data processing is categorized as research. These gaps create substantial ambiguity for collection of personal data and its AI training.

V. BEST PRACTICES ADAPTED BY OTHER COUNTRIES

- **Research Purpose:** Most data protection frameworks across jurisdictions do not explicitly exempt processing personal data for research. Instead, they treat research as a secondary use of data that does not require distinct lawful basis beyond the original basis for processing. Additionally, some laws allow for non-consensual data processing for research purposes, provided specific conditions and safeguards are met to protect individuals' rights. This approach balances enabling research activities while ensuring personal data is protected under established legal frameworks and ethical standards. In the European Union, the Article 89 of the GDPR [1viii]permits the secondary use of personal data for purposes such as archiving, statistical analysis, or scientific research, provided that appropriate safeguards are implemented to protect the rights of individuals. These safeguards typically involve technical and organizational measures that ensure data minimization, such as pseudonymization and limiting the scope of data processing to what is necessary for research purpose. This framework aims to balance enabling research while maintaining strong protections for personal data. In Japan, Article 76 of the Act on the Protection of Personal Information (APPI) [ix]provides an exemption from the requirement of obtaining data subject approval when personal data is collected secondarily and used in collaboration with an academic research institution. This exemption applies only if the processing is not exclusively for commercial gain and does not violate the rights and interests of the individuals concerned. In Singapore, the Personal Data Protection Act (PDPA) [x]allows organizations to use, collect, and disclose personal data for research purposes under specific conditions: (a) personally identifiable information is necessary for the research; (b) the research offers a clear public benefit; (c) the research outcomes will not be used for decisions that impact individuals; and (d) any published research results must not reveal individuals' identities.[xi]
- **Repurposing:** When compared to Section 6 of DPDP Act, GDPR's Article 5(1)(b) has the Compatibility test where the personal data collected shall only be for the specified explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes. Also, Article 5(1)(c) provides that personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed. GDPR sets a higher standard by requiring explicit compatibility checks and tightly limiting the scope of data collected. The DPDP Act adopts similar principles but may allow greater flexibility or less rigorous oversight in how necessity and purpose are applied. Article 23 of China's PIPL Act[xii], The party receiving personal information shall process personal information within the scope of the above purpose and method of processing and type of personal information. Where the party receiving personal information changes the original purpose and method of processing, it shall inform the individual and obtain his/her consent again in accordance with this Law. Thus, PIPL applies a more stringent and detailed approach to purpose limitation compared to DPDP. Article 15 of Japan's Act on the Protection of Personal Information (APPI) requires that a business

1

operator handling personal information specify the purpose of its use as clearly as possible. Any change in this purpose must remain within a scope reasonably related to the original purpose. Article 16 prohibits handling personal information beyond what is necessary to achieve the specified purpose without obtaining prior consent from the individual. This act enforces tighter control on how data repurposing occurs through a reasonableness lens and explicit notification rules.

- **Publicly Available Personal Data:** Article 9 of the GDPR states that Processing personal data that reveals racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, as well as processing genetic data, biometric data for unique identification, health data, or data concerning a person's sexual orientation, is generally prohibited. Under the GDPR, transparency and notification requirements apply even when personal data is publicly available, with strict obligations under Article 14 to inform individuals about the processing. Article 13 of China's Personal Information Protection Law (PIPL) broadly outlines the legal grounds for processing personal data. It permits processing personal data without obtaining consent from the individual if the data has been already disclosed by the individual or has been lawfully made public. However, such processing must be confined within reasonable limits, ensuring a proper balance between protecting the individual's rights and interests and serving the larger public interest. This framework seeks to enable appropriate use of publicly disclosed data while safeguarding personal rights. In Canada, under the Personal Information Protection and Electronic Documents Act (PIPEDA)[xiii], an exemption exists for the processing of publicly available personal information, but it applies only when the information falls within certain specific categories defined by the Regulations Specifying Publicly Available Information, SOR/2001-7 (December 13, 2000). The exemption is narrow and specific; not all publicly accessible data qualifies. For instance, certain types of information like voiceprints are excluded from the exemption. Organizations must still adhere to principles of limiting collection and use to the scope necessary for the intended purpose and respecting individuals' privacy.

VI. INTERNATIONAL CASE LAWS RELATING TO AI TRAINING

The P.M. et al. vs. OpenAI.[xiv]

This class action suit representing millions of internet users alleged that OpenAI violated multiple federal and state privacy laws, including the California Invasion of Privacy Act (CIPA). The core claim contends that OpenAI secretly scraped vast amounts of private information such as personal data, contact details, payment information, and private communications from the internet without users' consent or knowledge to train large language models like ChatGPT. The case challenges the notion that data posted online is freely available for commercial AI training, arguing that such collection amounts to an illegal interception and theft of personal information in violation of the right to privacy. Plaintiffs seek court orders requiring OpenAI to disclose the data it collected and to halt its non-consensual scraping practices, thereby shaping a legal framework to safeguard user privacy against opaque AI data harvesting.

Dinerstein v. Google, LLC, [xv]

A Dinerstein class action contends that through a sequence of corporate transactions enabling it to acquire and integrate an AI data-mining company named DeepMind and through partnerships with healthcare systems such as the University of Chicago, Google improperly obtained access to hundreds of thousands of patients' medical records in violation of the Health Insurance Portability and Accountability Act (HIPAA)[xvi]. Plaintiffs allege that Google used this illegally obtained personal health data to train machine-learning diagnostic and search algorithms, which Google then seeks to patent and monetize via fee-for-service, subscription, or standalone offerings. The Dinerstein complaint asserts multiple claims against the University of Chicago. The court dismissed the case holding that the plaintiffs' claim of potential future harm were deemed speculative and not an actual or imminent injury sufficient for monetary damages but sufficient for a claim of injunction.

Mutnick v. Clearview AI, et al.,[xvii]

This class action stems from Clearview AI's creation of a facial recognition database comprising millions of Americans, built from more than 3 billion photos Clearview scraped from online social media and other internet platforms. The plaintiff, David Mutnick, alleges that Clearview's AI facial recognition database has been sold to over 600 law enforcement agencies and other private entities to biometrically identify individuals who were unaware of and did not consent to Clearview's capture and use of their biometric data. Beyond monetary damages under BIPA (Illinois Biometric Information Privacy) Act, the plaintiff has filed a motion for a preliminary injunction to halt any further dissemination or use of the biometric data and to compel Clearview to implement stronger security measures to protect the database from additional breaches.

Burke v. Clearview AI*, Inc.*[xviii],

This case alleged that Clearview's collection and use of biometric data from consumers without adequate notice or consent violates privacy protections and supports the company's alleged use of such data to train facial recognition algorithms. The case frames potential violations of the California Consumer Privacy Act (CCPA) [xix]as unfair competition under California's UCL, arguing that Clearview's practices not only breach statutory requirements but also constitute deceptive and unlawful business conduct. Central to the claims is the contention that Clearview's data harvesting for AI training jeopardizes individual privacy and exposes consumers to biometric surveillance without meaningful informed consent, raising questions about regulatory gaps, consent standards, and the balance between innovative AI development and privacy rights. In addition to the fine of $9.4 million (£7.5 million), the ICO also ordered Clearview to delete data of U.K. residents from its system and to stop collecting and using the personal data of U.K. residents available online. The U.K.'s Information Commissioner said Clearview AI's practices are "unacceptable," not only identifying people but effectively monitoring their behaviour and offering it as a commercial service.

VII. SUGGESTIONS

- Transparency and data control measures should be tiered according to the level of risk associated with personal data and AI applications, with higher-risk categories subject to stricter safeguards. Privacy notices must not only disclose what data is collected but also explain what the AI can infer about users and how those inferences may affect them. Organizations should be bound by rule-based default expiration periods for data processing and must renew user consent whenever a new feature is introduced or the data purpose changes substantially. They must maintain consent records containing user ID, scope, timestamp, and policy version, and log how that consent is honoured in all data processing and exports. Each data flow should carry attached policy decisions to enable an immutable audit trail, and users must have a clear, accessible way to opt out of personalization or automated decisions based on inferred data.
- Clarify and narrow the research exemption under Section 17(2)(b) to prevent its misuse for commercial AI development disguised as research. Instead of the broad exemption allowed under the DPDP Act, adopt a more nuanced approach aligned with Article 89 of the GDPR, which mandates appropriate safeguards such as data minimization, purpose limitation, and pseudonymization for research activities. The law should clearly define the criteria for invoking the research exemption and establish independent oversight mechanisms to ensure that this provision is not exploited to bypass user consent or compromise data protection principles.
- Strengthen purpose limitation and repurposing controls under Section 6 by introducing an explicit compatibility test, similar to Article 5(1)(b) of the GDPR, to ensure that any further processing of personal data is restricted to purposes compatible with the original collection intent. Require data fiduciaries to notify users and obtain fresh consent whenever data initially collected for one purpose, such as service improvement, is repurposed for commercial AI model training. Drawing from Japan's APPI Articles 15 and 16, mandate that any change in purpose must meet standards of reasonableness and be supported by informed user consent to uphold transparency and prevent misuse.
- Regulate the use of publicly available data more stringently by closing the broad exemptions under the DPDP Act that allow unrestricted processing of such data. Require transparency obligations and protection of contextual integrity, drawing from GDPR Article 14 and China's PIPL Article 13, to ensure that the original context of disclosure is respected. Explicitly limit the assumption that publicly available data grants an unlimited license for AI training by distinguishing between personal disclosure and commercial exploitation. Establish clear individual rights to restrict or object to the use of their publicly available data for AI model training to safeguard autonomy and prevent misuse.
- Enhance informed consent mechanisms to address the specific challenges posed by AI systems by developing consent frameworks that are explicit, granular, and truly informed, moving beyond the broad and opaque Terms of Service often used today. Legislation should require data fiduciaries to provide distinct consent options for different processing purposes, including core service operation, analytics and product improvement, personalization, advertising, AI model training and evaluation, and data sharing with partners. This approach ensures that individuals retain meaningful control over their data and understand the implications of each consent choice.

VIII. CONCLUSION

The central conclusion of this paper is that the phenomenal growth of powerful Artificial Intelligence (AI) models in India is built upon a precarious foundation: the silent, non-transparent harvesting of vast amounts of personal data. Our analysis confirms that the deliberate obscurity in data accumulation processes are actively exploited to feed AI training algorithms, raising profound concerns about privacy infringement. This paper demonstrates that while the DPDP Act, 2023 establishes a foundation for data protection, its broad research exemptions, vague purpose limitation rules, and liberal approach to publicly available data leave critical regulatory gaps. Building on global best practices, it is essential to introduce stricter tiered transparency requirements for high-risk applications, require granular informed consent, and narrow the research exemption to prevent commercial misuse. Controls over data repurposing and the use of publicly available data must be enhanced to ensure alignment with contextual integrity and user expectations. By mandating these reforms, and adopting AI-focused governance rooted in accountability, transparency, and reasonableness, India can safeguard user rights and ensure that responsible AI development upholds dignity, trust, and autonomy for all.

IX. REFERENCES

[i] https://www.rws.com/artificial-intelligence/train-ai-data-services/blog/how-ai-is-trained-the-critical-role-of-ai-training-data/#:~:text=AI%20training%20data%20is%20a,process%20and%20generate%20human%20language.

[ii] https://share.google/PgRriCUe9YQ0vvlyX

[iii] https://www.postman.com/what-is-an-api/

[iv] https://www.protecto.ai/blog/impact-of-ai-on-user-consent-and-data-collection

[v] https://www.gatewayhouse.in/lessons-from-the-chatgpt-confession-files/

[vi] Digital Personal Data Protection Act, 2023 (DPDP Act) (No.22 of 2023)
[vii]https://chambersofnamratapahwa.com/strengthening-data-protection-in-india-the-digital-personal-data-protection-act-2023-and-ai/

[viii] General Data Protection Regulation – 2016/679
[ix] Act on the Protection of Personal Information (Act No. 57 of 2003)
[x] Personal Data Protection Act , 2012
[xi] https://fpf.org/blog/five-ways-in-which-the-dpdpa-could-shape-the-development-of-ai-in-india/

[xii] Personal Information Protection Law of the People's Republic of China, 2021
[xiii] Personal Information Protection and Electronic Documents Act (PIPEDA), 2012
[xiv] *Case No. 3:23-cv-03199 (N.D. Cal.)*
[xv] *20-3134(7th Cir.2023)*

[xvi] Health Insurance Portability and Accountability Act (HIPAA), 1996
[xvii] *1:20-cv-00512 (N.D. Ill.) (filed January 22, 2020).*

[xviii] *3:20-cv-00370-BAS-MSB (S.D. Cal.) (filed February 27, 2020).*

[xix] California Consumer Privacy Act , 1967