The Legal Status of Artificial Intelligence (AI) Memory: Can Individuals Demand the Deletion of Information Learned by Generative AI?

Bhavi Pungliya[1]

ABSTRACT

Once information relating to a person's private life is reproduced by an artificial intelligence system, the person can request its immediate deletion. Section 12 of the Digital Personal Data Protection (DPDP) Act, 2023 recognizes the right to erasure. However, traditional erasure affects the records stored in the database and does not directly affect the personal information contained in the trained AI model. This article considers whether India's legal right to erasure makes sense when personal information is part of the training data, corpus fine-tuning, integration, or search architecture of generative artificial intelligence systems. Neither the DPDP Act nor the 2025 DPDP Rules specifically address this issue in the context of training AI models. Adopting a doctrinal approach, the article reviews the structure of the DPDP, constitutional jurisprudence on privacy, comparative developments in the General Data Protection Regulation (GDPR), Indian jurisprudence on the right to be forgotten, and the technical literature on machine learning and unlearning. The author argues that deleting the original record does not necessarily destroy the information the model has captured, and the model can continue to reproduce the information. This article proposes, a step-by-step framework based on identifiability, reproducibility, harm, and technical feasibility, supplemented by guarantees at the level of results and access, when complete model erasure is not possible. It places the burden on the data fiduciary to prove technical impossibility and asserts a significantly reduced duty.

KEYWORDS

Generative AI; Data Erasure; AI Memory; Digital Personal Data Protection Act; Machine Unlearning; Informational Autonomy; Right to be Forgotten

[1] PhD Scholar, Sri Aurbindo Institute of Law, Indore, Madhya Pradesh, India.

INTRODUCTION

A person uses a widely available generative artificial intelligence assistant and discovers, quite by accident, that it can recount accurate details about her medical history, her home address, and a period of unemployment she had never made public. She traces the likely source to a dataset the AI developer once used and asks for the information to be deleted. The company duly deletes the underlying record from its database and confirms the request has been completed. Weeks later she asks the same assistant, in a different conversation, about herself, and it still produces an approximation of the same private facts.

Has her information actually been deleted? The record is gone, yet the assistant still knows something. This is the puzzle this article investigates. It asks a narrow and specific question: once personal information has passed through the training, fine-tuning, or retrieval architecture of a generative AI system, what does a legal right to deletion actually require, and can Indian law compel it in any meaningful sense.

The Digital Personal Data Protection Act, 2023[2] provides for a right to correction and erasure, although the relevant provisions are not yet in force. This article argues that treating the deletion of a source record as equivalent to the deletion of its influence on a trained model is a category error, one that produces symbolic compliance without substantive protection. It examines the Indian constitutional and statutory position, draws lessons from the European Union's (EU) more developed but still unresolved experience under the General Data Protection Regulation (GDPR), and proposes a graduated framework that ties the erasure obligation to identifiability, reproducibility, and demonstrable harm, while shifting the burden of proving technical infeasibility onto the entity that trained the model.

UNDERSTANDING WHAT AI MEMORY MEANS

A generative AI system is not the same as a filing cabinet. At this juncture, it is vital not to exaggerate its memory power. In the training process, the model consumes a large amount of text or other type of data and adjusts its millions or billions of internal numerical values, its parameters or weights, which means that it can better predict what comes next in the sequence. For most individual pieces of training data, no identifiable copy survives this process in any

[2] Digital Personal Data Protection Act, No. 22 of 2023 (India).

conventional sense; the influence of a single record is diffused across the model's parameters alongside the influence of countless other records.

Memorisation is the exception to this general picture, and it is well documented rather than speculative. Researchers have shown that large language models can, under the right prompting conditions, reproduce specific sequences from their training data verbatim, including personal information, and that larger models memorise more than smaller ones.[3] A separate study demonstrated that even production systems with safety alignment can be induced to emit training data through targeted prompting strategies.[4] This means some information genuinely persists in an extractable form, while the greater part exists only as diffuse statistical influence that shapes outputs through inference rather than recall.

Four further categories must be kept distinct. Fine-tuning uses a smaller, often more sensitive dataset to adjust a pretrained model's behaviour, and the resulting adjustments are typically easier to trace and reverse than the base model's original training. Retrieval-augmented generation architectures store documents in an external vector database that the model consults at query time; deleting an entry from that database is technically straightforward and closely resembles deleting a conventional record, because the information never entered the model's parameters at all. Conversational memory, the history an application retains of past exchanges with a particular user, is again an ordinary storage layer external to the trained model itself. None of these three should be confused with the fourth category, namely information that has actually shaped the model's parameters during pretraining or fine-tuning, which is the layer that resists conventional deletion. The National Institute of Standards and Technology's Generative AI Profile treats leakage through memorisation, inference, and extraction as a distinct data privacy risk precisely because it does not behave like ordinary data storage.[5]

THE INDIAN LEGAL FRAMEWORK

The constitutional starting point is *Justice K.S. Puttaswamy (Retd) v. Union of India*[6], in which nine judges concluded that privacy is protected by Articles 14, 19 and 21[7] of the Constitution

[3] Nicholas Carlini et al., *Extracting Training Data from Large Language Models*, 30th USENIX Sec. Symposium 2633 (2021).
[4] Milad Nasr et al., Scalable Extraction of Training Data from (Production) Language Models, arXiv:2311.17035 (Nov. 28, 2023).
[5] Nat'l Inst. of Standards & Tech., Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1 (2024).
[6] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1 (India).
[7] INDIA CONST. art. 14, 19, 21.

and that informational privacy, understood as the capability of an individual to manage information about oneself, is part of this right. Although the case did not concern artificial intelligence, its reasoning provides a constitutional basis for assessing whether an individual retains meaningful control over personal information after it has been processed by a computing platform. At the statutory level, Section 12 of the Digital Personal Data Protection Act, 2023 grants a data principal the right to correction, completion, updating, and erasure of personal data that she has previously consented to being processed, and requires the data fiduciary to notify any associated data processor of the erasure.[8] Two limits are important. First, the right is tied to consent based processing rather than operating as a free standing entitlement against any use of one's information, so data drawn from legitimate uses recognised elsewhere in the Act, or from information the individual has herself made public, sits outside its immediate reach. Second, and more significantly for this article, the Act contains no provision addressing what erasure means once information has been used to train a model rather than merely stored as a record. The drafters were legislating against the template of conventional databases.

The Digital Personal Data Protection Rules, 2025 were notified on 13 November 2025, giving operational effect to the Act, but their commencement is staggered across three dates, with substantive obligations, including the detailed mechanics of correction and erasure requests, scheduled to come fully into force only on 14 May 2027.[9] As of the date of writing, the Data Protection Board of India exists and can receive complaints, yet the erasure machinery that would need to grapple with AI training data has not yet become fully operative, leaving the central question of this article formally unresolved in Indian subordinate legislation.

Indian courts have separately developed a right to be forgotten as a facet of Article 21[10], though in a different technical context. In *Laksh Vir Singh Yadav v. Union of India*, the Delhi High Court held that individuals may seek de-indexing and masking of personal information from name based searches on court record databases and search engines, balancing informational privacy against open justice.[11] The Madras High Court had earlier reached a similar conclusion in *Karthick Theodore v. Registrar General*, though the Supreme Court stayed that direction pending further consideration (*Ikanoon Software Development Pvt. Ltd. v. Karthick Theodore*

[8] Digital Personal Data Protection Act, No. 22 of 2023, § 12 (India).
[9] Digital Personal Data Protection Rules 2025, GSR 846.
[10] INDIA CONST. art. 21.
[11] Laksh Vir Singh Yadav v. Union of India & Ors., 2026:DHC:4891 (Delhi H.C. May 29, 2026).

and Ors.)[12]. These cases are instructive analogues rather than direct authority on AI memory: they concern suppressing the retrieval of an existing, unaltered record, which is technically closer to deleting a retrieval-layer entry than to altering a trained model's parameters, and they confirm that Indian courts are willing to fashion layered remedies, such as de-indexing short of destruction, when full erasure is impractical or undesirable.

IS DELETION LEGALLY POSSIBLE AFTER MODEL TRAINING?

The central difficulty is that several operations are routinely described using the single word deletion when they are, in fact, distinct events with different feasibility profiles. Deleting a source record from a database is straightforward and is what the DPDP Act's erasure right naturally reaches. Preventing future access to information, for instance by removing an entry from a retrieval database or blocking a cached output, is also generally feasible because the information sits outside the model's parameters. Preventing reproduction of information that has already shaped a model's weights is considerably harder, since it usually requires either output level filtering, which suppresses symptoms without touching the underlying representation, or some form of retraining.

Full retraining without the objectionable data would satisfy the underlying concern most completely, but it is expensive to the point of impracticality for large foundation models, given the scale of compute involved in a single training run.[13] Machine unlearning has emerged as an intermediate technical response, seeking to approximate the effect of never having trained on a given example without the cost of full retraining. The literature is candid that most unlearning methods are approximate rather than exact, that verifying success is itself an open research problem, and that the non-deterministic character of the training process makes it difficult to establish with confidence which weights were influenced by which input in the first place.[14] Recent scalable extraction results further show that alignment techniques, the safety layers applied after pretraining, reduce but do not eliminate the risk that a sufficiently determined party can extract memorised material from a deployed model.[15]

[12]iKanoon Software Development Pvt. Ltd. v. Karthick Theodore & Ors., SLP (C) No. 15311/2024 (India July 24, 2024).
[13]Henrik Nolte et al., Machine Learners Should Acknowledge the Legal Implications of Large Language Models as Personal Data, arXiv:2503.01630 (Mar. 3, 2025).
[14]Bjørn Aslak Juliussen et al., Algorithms That Forget: Machine Unlearning and the Right to Erasure, 51 Comput. L. & Sec. Rev. 105885 (2023).
[15]Nasr et al., *supra* note 4.

The legally significant conclusion is that deleting the original database entry and eliminating the model's capacity to reproduce equivalent information are not the same act, and current Indian law addresses only the former with any clarity. A data fiduciary that erases a record and reports compliance may be accurately describing what it has done to its database while leaving open, and typically unaddressed, the separate question of whether the same information can still be elicited from the model it trained.

LESSONS FROM COMPARATIVE LAW

The European Union offers the most developed comparative experience, though not a solved problem. Article 17 of the GDPR grants a right to erasure but, as commentators have observed, gives no technical guidance on what erasure means for a machine learning model trained on the data in question.[16] The GDPR does not prescribe a specific technical method for erasing information from a machine-learning model. Article 17(2) expressly requires reasonable steps, taking account of available technology and the cost of implementation, where personal data has been made public, while Article 17(3) provides specified exceptions to the right to erasure.[17] The Court of Justice's earlier decision in *Google Spain*, which grounded a right to de-listing search results rather than requiring destruction of the underlying material, prefigured this preference for layered, retrieval-level remedies over root and branch deletion, and its logic maps more comfortably onto retrieval-augmented systems than onto trained parameters.[18]

More recent European developments reinforce the shift away from treating erasure as the primary tool. The EU AI Act imposes transparency obligations on providers of general purpose AI models, requiring training data summaries and technical documentation from August 2025, which indirectly assist data subjects in establishing whether their information was used without creating a direct erasure remedy of its own.[19] The European Commission's 2026 Digital Omnibus proposal goes further, proposing to let AI developers rely on legitimate interest as a lawful basis for training, coupled with an unconditional opt-out, rather than requiring case by case erasure after the fact.[20] Read together, these developments suggest that European policy is moving toward front-loaded safeguards, transparency and opt-out at the point of training,

[16]Juliussen et al., *supra* note 13.
[17]Regulation (EU) 2016/679, art. 17(2)-(3), 2016 O.J. (L 119) 1.
[18]Google Spain SL v. Agencia Española de Protección de Datos (AEPD).
[19]Regulation (EU) 2024/1689 (EU AI Act), arts. 53 to 55.
[20]International Association of Privacy Professionals, ''EU Digital Omnibus Amendments to GDPR to Facilitate AI Training Miss the Mark'' (28 May 2026).

precisely because back-end erasure from an already trained model has proved so difficult to operationalise. India, still at the design stage of its subordinate legislation, is well placed to build such front-loaded safeguards into the DPDP Rules rather than replicate the GDPR's after the fact struggle.

THE PROBLEM OF PROOF AND TECHNICAL VERIFICATION

An ordinary erasure request can be verified because the individual, or an auditor, can point to the specific record and confirm it is gone. An AI training request cannot be verified in the same way, because a data principal has no means of inspecting a model's parameters to determine whether her information influenced them, and even the developer may not be able to state with certainty which examples shaped which outputs, given the non-deterministic and distributed nature of the training process. Requiring the data principal to prove that her information persists within a system she cannot access would render the right illusory.

The more workable allocation places the burden on the data fiduciary. Once an individual makes a plausible showing that her personal data was processed by the fiduciary and that comparable information can be elicited from the resulting system, whether through direct prompting or documented extraction techniques, the fiduciary should bear the burden of demonstrating either that the information cannot in fact be reproduced or that mitigation short of full erasure has been applied. This allocation tracks the general principle that the party in control of the technical architecture, not the individual outside it, should bear the burden proportional to that control.

The obvious objection is trade secrecy: AI developers resist disclosing training data composition and model architecture as proprietary and commercially sensitive. This tension need not be resolved through public disclosure. The EU AI Act's model, in which providers submit technical documentation to a regulator rather than the public, offers a workable template.[21] India's Data Protection Board could perform an analogous confidential verification function, receiving evidence of training data provenance and mitigation measures without requiring public disclosure of proprietary architecture, so that verification and confidentiality are not treated as mutually exclusive**.**

[21]Regulation (EU) 2024/1689, art. 53, 2024 O.J. (L 1689) 1.

TOWARDS AN INDIAN FRAMEWORK FOR AI DATA ERASURE

Rather than proposing an unqualified duty to delete information from a trained model, a duty that would be either unenforceable or crippling to AI development if taken literally, this article proposes a tiered framework calibrated to risk and feasibility.

First, source level erasure should always be required and is already achievable: the original record, any fine-tuning dataset entry, and any retrieval database entry must be deleted on request, since none of these operations depends on altering trained parameters. Second, where the data principal or the Board can demonstrate, through documented extraction or membership inference style testing, that the information remains reproducible from the model itself, the fiduciary should be required to apply output level mitigation, such as filtering or refusal behaviour targeted at the specific information, pending a more durable remedy. Third, for sensitive personal data shown to be both reproducible and capable of causing material harm, such as health, financial, or identity information, the fiduciary should be required to apply machine unlearning or equivalent parameter level mitigation at the next scheduled retraining or fine-tuning cycle, verified by the Board rather than by the individual, since on demand exact unlearning is not currently a reliable technology. Fourth, full retraining or withdrawal of a model should remain a reserved remedy for cases of severe, repeated, or wilful non-compliance, reflecting its disproportionate cost relative to most individual grievances. Underlying all four tiers, the burden of establishing infeasibility rests on the fiduciary, and the Board should be empowered to conduct confidential technical audits without requiring public disclosure of proprietary information.

This framework departs from a simple transplantation of the GDPR's right to erasure by making explicit what the DPDP Act currently leaves unstated: that different loci of persistence call for different remedies, and that the individual's practical entitlement should not depend on winning an argument about model internals she has no means of examining.

CONCLUSION

Section 12 of the Digital Personal Data Protection Act, 2023, once brought into force, will provide, not for a world of trained parameters, and applying it literally to a generative AI system produces an answer that satisfies the letter of the statute while leaving the underlying informational autonomy interest recognised in *Puttaswamy* substantially unprotected. Deleting

a database entry is necessary but not sufficient once information has shaped a model capable of reproducing it. The honest answer to this article's opening hypothetical is that the woman's information was only partially deleted: the record was destroyed, but the system's capacity to reproduce equivalent information may well have survived that deletion untouched.

India should not respond by importing an unworkable absolute obligation to erase information from trained models, a standard even the more mature European framework has struggled to enforce. It should instead build, through amendment of the DPDP Rules before their substantive provisions take full effect in May 2027, a graduated framework that requires source level deletion as a baseline, escalates to output filtering and scheduled parametric mitigation where reproducibility and harm are demonstrated, and places the burden of proving technical impossibility on the data fiduciary rather than the individual. Only such a framework can give the statutory right to erasure genuine content in an age in which personal information no longer merely sits in a database, but may live on, in diffuse and technically stubborn form, inside the memory of a machine.