The Indian Journal for Research in Law and Management
Open Access Law Journal – Copyright © 2026
Editor-in-Chief – Dr. Muktai Deb Chavan; Publisher – Alden Vas; ISSN: 2583-9896

This is an Open Access article distributed under the terms of the Creative Commons Attribution- Non-Commercial-Share Alike 4.0 International (CC-BY-NC-SA 4.0) License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium provided the original work is properly cited.

AI TRAINING DATA AND PRIVACY: CAN PERSONAL DATA BE USED TO TRAIN AI?

Every time you post a photo, write a review, or publish a blog, you leave a small piece of yourself on the internet. Most of us never imagined those pieces would one day be collected, in bulk, to teach machines how to write, speak, and think. Yet that is how many of today's AI systems are built. This raises a question that lawyers, regulators, and ordinary users are all asking: is it legal to use personal data to train AI?

WHY TRAINING DATA MATTERS

Generative AI models learn by finding patterns in enormous datasets, often scraped from websites, social media, forums, and digitised books. These datasets rarely contain only neutral facts. They also hold names, addresses, opinions, photographs, medical discussions, and private conversations that were never meant to be anyone's raw material.

The privacy concern is not just that the data is collected. Models can sometimes memorise and reproduce fragments of what they were trained on, including personal details. Once information is absorbed into a model, it is very hard to locate and pull back out.

THE INDIAN POSITION

India's Digital Personal Data Protection Act, 2023 (DPDP Act) is built around consent and purpose limitation. A company that processes personal data must generally have the person's consent or fall under one of the Act's "legitimate uses". The consent must be free, specific, informed, and tied to a stated purpose.

Here is the difficulty. A person who shared a post in 2015 consented, at most, to that platform displaying it. They did not consent to it becoming training material for a commercial model a decade later. The purpose has changed, and the original consent does not clearly stretch that far.

The Indian Journal for Research in Law and Management
Open Access Law Journal – Copyright © 2026
Editor-in-Chief – Dr. Muktai Deb Chavan; Publisher – Alden Vas; ISSN: 2583-9896

This is an Open Access article distributed under the terms of the Creative Commons Attribution- Non-Commercial-Share Alike 4.0 International (CC-BY-NC-SA 4.0) License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium provided the original work is properly cited.

The Act, however, contains an exemption that matters a great deal for AI developers. It does not apply to personal data that the individual themselves made publicly available. AI companies can argue that scraping such content falls outside the Act. Critics respond that "publicly available" is not the same as "free to use for any purpose", and that people posting online never anticipated large-scale commercial training. The statute does not directly address AI training, and the question remains open for courts and regulators to settle.

The constitutional backdrop also matters. In *Justice K.S. Puttaswamy v. Union of India* (2017), the Supreme Court recognised privacy as a fundamental right. Any law or practice that affects informational privacy has to be understood against that foundation, even when the statute itself is silent on a new technology.

THE GLOBAL PICTURE

Europe has been more direct. Under the GDPR, developers need a lawful basis for processing personal data, and many have relied on "legitimate interests". Regulators have not accepted this blindly. Italy's data protection authority temporarily blocked ChatGPT in 2023 over concerns about the lack of a legal basis and inadequate transparency, and later fined OpenAI for privacy violations. Other European regulators have examined how companies like Meta and X use public posts for training.

The lesson from these actions is that "it was online, so it was fair game" is not a safe legal argument. Transparency, a defensible legal basis, and a workable way for people to object are becoming the expected minimum.

THE HARDEST PROBLEM: ERASURE

Modern privacy laws give people the right to correct or erase their data, and the DPDP Act includes this right too. But how do you delete someone from a trained model? Removing a record from a database is simple. Removing its influence from billions of learned parameters is not. The emerging field of "machine unlearning" is trying to solve this, but it is far from a settled technical answer.

The Indian Journal for Research in Law and Management
Open Access Law Journal – Copyright © 2026
Editor-in-Chief – Dr. Muktai Deb Chavan; Publisher – Alden Vas; ISSN: 2583-9896

This is an Open Access article distributed under the terms of the Creative Commons Attribution- Non-Commercial-Share Alike 4.0 International (CC-BY-NC-SA 4.0) License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium provided the original work is properly cited.

So the law promises a right that the technology cannot yet reliably deliver. Either regulators will push developers to build better tools, or the right of erasure will be quietly weakened in practice. Neither outcome is comfortable.

FINDING A BALANCE

It would be unrealistic to argue that AI should never touch personal data. Useful systems in healthcare, language translation, and accessibility depend on learning from real human information. But it would be equally unwise to let innovation run ahead of people's basic dignity and control over their own information.

A reasonable path forward might include:

- **Data minimisation:** collect only what is actually needed, and filter out sensitive personal information before training.
- **Transparency:** tell people clearly when and how their data may be used.
- **Meaningful opt-outs:** give individuals a real, easy way to refuse.
- **Privacy-enhancing techniques:** use anonymisation, differential privacy, and synthetic data wherever possible.
- **Clear regulation:** update the law, or issue guidance, so that AI training is addressed explicitly rather than left to interpretation.

CONCLUSION

So, can personal data be used to train AI? The honest answer is: sometimes, but not freely and not without responsibility. The law, in India and elsewhere, is still catching up with a technology that moved faster than anyone predicted. Until it does, the safest approach is to treat personal data not as a free resource to be mined, but as something that belongs to a person with rights, expectations, and a reasonable claim to be asked first.

Technology will keep advancing. Whether trust advances with it depends on the choices that developers, lawmakers, and the rest of us make now.