# Data Empowerment and Protection Architecture (DEPA) for AI Training:

## *A Discussion Paper*




**Naavi**

(Vijayashankar Nagaraja Rao)

Revised edition | 26 November 2026

*In the light of AIGSI Version 3.0*

## Data Empowerment and Protection Architecture (DEPA) for AI Training: A Discussion Paper

Naavi (Vijayashankar Nagaraja Rao)

Revised edition, 26 November 2026, in the light of AIGSI Version 3.0

(First edition presented at IDPS 2026, Bengaluru, 21 November 2026)

> **About this revised edition.** The first edition of this paper, presented at IDPS 2026 in Bengaluru, set out how AIGSI Version 2.0 could fill the governance gaps in DEPA for AI Training. Since then FDPPI has prepared AIGSI Version 3.0 as a discussion paper for its proposed Special Interest Group (SIG) on AI Governance. Version 3.0 deals directly with several issues this paper raises: where the liability chain ends, what happens when roles blur, who becomes a Model Owner when a model is fine-tuned, fairness and collective harm, copyright, release controls, and a single audit that produces a CERT-In opinion, a DPDPA opinion and an IT Act opinion.
>
> This edition rewrites Part E around AIGSI Version 3.0, adds an AIGSI column to the table of open issues in Part C, adds cross-references in Part D, adds recommendations to Part F, and adds Questions 36 to 41 to the Annex. **Questions 1 to 35 keep their original numbers**, so responses already sent quoting those numbers remain valid. The analysis of the DEPA handbook in Parts A to D is otherwise unchanged.

### Preface

On 31 October 2026, MeitY released the *Handbook on Data Empowerment and Protection Architecture (DEPA) for AI Training* at a roundtable in Bengaluru. Several dignitaries spoke at the launch.

**Shri S. Krishnan, Secretary, MeitY:**

> "India's AI ambitions must be matched by the ability to use data responsibly and securely. The DEPA Training Framework enables trusted data collaboration at scale, unlocking access to large-scale datasets for AI training while safeguarding privacy, commercial confidentiality and institutional control. MeitY is fully committed to taking this framework forward to advance AI innovation in the country."

**Prof. Ajay Kumar Sood, Principal Scientific Adviser to the Government of India:**

> "Under DEPA, the model travels to the data, rather than the data travelling to the model. This enables trusted, multi-party and cross-border data collaboration, accelerating scientific discovery and helping solve problems that no single institution can address alone. DEPA is India's contribution to global science diplomacy, advancing the efforts of institutions such as the World Bank and the United Nations to shape a more inclusive and responsible global AI ecosystem."

**Shri Sharad Sharma, Co-founder, iSPIRT Foundation:**

> "Today marks the whole data economy (of many pillars) coming into place. This is easily a more significant moment since February 2016, when the Unified Payment Interface (UPI) was announced. India has the most distinctive, safe and flexible data economy in the world. Because of this, India now has a place in the AI stack like never before."

**Shri Anjani Rathor, Group Head – Digital Banking and Customer Experience, HDFC Bank:**

> "Trusted data collaboration can unlock significant value for AI in banking. Our experience with DEPA Inference has demonstrated the promise of secure data collaboration. We see a compelling opportunity to build on this experience through DEPA Training."

**Shri Kris Gopalakrishnan, Co-founder, Infosys and Chairman, Axilor Ventures:**

> "India has an opportunity to lead in safe AI by enabling trusted access to data for research and innovation. DEPA Training can help researchers harness diverse datasets to develop more robust AI models, improve collaboration across multiple teams, and advance scientific discovery, while safeguarding the rights and interests of data custodians and contributors."

## Why this paper

The 41-page handbook acknowledges that the initiative was pioneered by the iSPIRT Foundation, a body of industry leaders who earlier ran product initiatives within NASSCOM. iSPIRT has now proposed a new DEPA architecture that lets AI model trainers access bulk data from Indian companies and the Government. The framework is referred to as DEPA-AIT (DEPA for AI Training).

In effect, the framework opens access for AI model developers to valuable personal data, which is a sovereign Indian asset. Access is not reserved for Indian developers. So this is not an initiative to build indigenous AI capability. It may instead end up enriching OpenAI, Anthropic, Google, Meta and DeepSeek with data from the Indian Government and Indian companies.

The framework will create a facility where data from different sources in India is collected and made available for AI model training. Prudent Indian companies may hesitate to share their data on such a platform, but the Government may well do so. DEPA-AIT may therefore become mainly a channel for moving Government-held personal data of Indian citizens to the leading AI model developers in the US.

The Secretary, MeitY has invited "actionable recommendations" for refining and co-creating this framework, which is presented as a "Digital Public Infrastructure" for the benefit of all. This paper responds to that invitation.

Naavi presented the first edition of this discussion paper at IDPS 2026, Bengaluru on 21 November 2026, and presents this revised edition at IDPS 2026, Hyderabad on 28 November 2026, to invite comments from fellow professionals.

Comments may be sent to naavi@fdppi.in with the subject line "DEPA for AI Training". Participants are requested to respond to the questions in the Annex: Questions for Discussion at the end of this paper, quoting the question number.

**Naavi**

## Part A: The core concerns

### The DPDPA context

The backdrop to this document is the DPDPA 2023, whose main obligations come fully into force in about six months. Under the Act, a Data Fiduciary may process personal data only with the specific consent of the Data Principal, unless a recognised exception applies. Consent is purpose-specific. The same data cannot be used for another purpose, including selling or otherwise sharing it for AI model training.

The Act permits a few uses of personal data without consent as "legitimate uses" under Section 7. Separately, Section 17(2)(b) exempts processing for research, archiving or statistical purposes, subject to prescribed conditions. Data that is genuinely anonymised falls outside the Act altogether.

Any other use of personal data, whether direct or through a processor, is prohibited and could lead to penalties that may add up to around ₹850 crore in total. Breaches may also result in imprisonment of company officials for their negligence and in compensation payable to affected Data Principals.

Given this exposure, the Government knows that no private company can easily be persuaded to share its data for AI training. The Government itself may share data, but it too risks liability under DPDPA 2023.

### The "Data Deadlock" and the pressure to break it

The Government has therefore recognised a "Data Deadlock": rich data sits with industry and public bodies but cannot be used for AI training. Industry is now pressing the Government to break this deadlock by creating a legal channel through which the personal data of Indian citizens can reach AI model developers.

In Naavi's view, DEPA-AIT is a framework that facilitates such personal data transport.

DPDPA 2023 is already before the Supreme Court on questions of constitutional validity. At this time, it is surprising that the Government is yielding to commercial pressure to dilute the Act through manoeuvres such as DEPA for AI Training. If this is raised before the Court as evidence of the Government's bad faith in protecting privacy, it could seriously weaken the Government's defence in the cases now pending.

### A "UPI for data", built without visible consultation

The vision is to make this framework a Digital Public Infrastructure like UPI, where organisations contribute data and AI model developers harness it.

The document says it was developed through extensive collaboration between Government, industry and academia. But no details of such consultation are publicly available. Since this is a public project, the record of how it was developed should be obtainable through RTI from the public institutions involved.

### The CIBIL precedent

The possibility that this framework is a backdoor through which AI industry can access and exfiltrate the personal data of 140 crore Indians is glaring. A similar data laundering exercise occurred around 2014, when TransUnion was allowed to take control of CIBIL through a transfer of controlling shares. Control over some 500 million data records worth billions of dollars passed with it, at a throwaway price of around ₹3,800 crore. The current proposal must therefore be viewed with suspicion until proved otherwise.

### Endorsements and the "need to share" narrative

The document carries endorsements from several dignitaries, in an effort to build support for highlighting the "need to share data" for AI innovation. As one foreword puts it, the handbook is "a statement of intent and a positive contribution to the global thinking on data and AI governance."

If the Government's intention were national benefit, it would have recognised that this proposal places a treasure of sovereign data in the hands of international AI companies, from whom Indians will later have to buy services back. It is like the days of the British, when raw materials were shipped from India to England and value-added products imported back.

It is impossible to believe that MeitY, iSPIRT or NASSCOM is unaware of this. The inference is that the Government is yielding to the commercial pressure of the AI lobby led by Meta, Google, Anthropic and OpenAI.

The Comptroller and Auditor General (CAG) recently flagged the need to determine the fair value of data held by PSUs, and instructed auditors to have it recognised during audits. It is therefore surprising that MeitY now proposes to give away a national asset without proper thought.

Had sharing been restricted to Indian companies, there would have been a defensible logic that Indian data enriches Indian AI models. No such restriction is proposed. As drafted, this appears to be a scheme for making Indian data available to foreign AI model developers.

The proposal needs to be understood and debated in this context.

## Part B: What the framework proposes

### Principles and participants

The framework is said to rest on the ADEPTS principles (Accountability, Democratization, Evolvability, Privacy, Transparency and Security), which are meant to establish the foundation of trust for India's data ecosystem.

The two principal participants are:

- **Training Data Providers (TDPs):** institutions such as hospitals, banks and public agencies that hold valuable training data.
- **Training Data Consumers (TDCs):** innovators such as start-ups, researchers and enterprises.

The stated purpose is to bring TDPs and TDCs onto a common techno-legal platform, where secure, private and compliant data exchange and AI modelling can produce data-driven solutions for societal impact.

The framework rests on four core pillars:

1. Data Exchange Platform
2. Electronic Contracts
3. Secure Data Environment (SDE)
4. Differential Privacy

Through this combination of technology, contracts, privacy-enhancing techniques and institutional oversight, DEPA-AIT seeks to overcome the "Data Deadlock".

### The Data Exchange Platform: a data repository

The Data Exchange Platform (DEP) is expected to play the role of a data repository. TDPs provide bulk data, which is anonymised and encrypted and made available for download and use by registered TDCs and AI model builders. Two options are proposed. In the first, the TDP does the anonymisation. In the second, the DEP does it.

It is not clear whether the private sector will adopt the framework and absorb the DPDPA risks. Government departments, however, may be persuaded by MeitY to adopt it. DEPA-AIT then becomes the instrument through which data held by the Government, a sovereign data asset, reaches OpenAI, Anthropic, Meta or Google. It will be interesting to see how the CAG reacts to this proposition.

### The techno-legal approach to governance

DEPA-AIT proposes to bring all critical stakeholders (data providers, AI developers, government representatives, technology specialists and domain experts) into a consortium of market players that would provide oversight for a trustworthy and responsible innovation ecosystem.

It recognises the need for an oversight body to anchor essential functions across the AI development lifecycle:

- setting clear standards for privacy, ethics and security

- registering eligible participants
- overseeing techno-legal mechanisms for the data sharing procedure
- facilitating dispute resolution

### Integration with national data exchanges

The document states that DEPA becomes more powerful when combined with complementary national data exchanges: AI Kosh, API Setu, the NIC API Exchange Platform (NAPIX), the National Data and Analytics Platform (NDAP), the Geospatial Data Infrastructure (GDI) and Open Government Data (OGD). Secure access to these platforms, it says, would enrich AI models across all sectors with vital context.

This indicates an intention to integrate all these data sources into the platform. In Naavi's view, the intention is clear: to put all available personal data of citizens to use for the benefit of global players, so that India ends up eating out of their hands later.

The document does acknowledge the risks under DPDPA, but the acknowledgement lacks credibility.

### Why consent is set aside

The document argues that the earlier DEPA framework and similar architectures, based on consent between the Data Principal and the Data Fiduciary, cannot be followed for AI. It says a different model is needed for sharing personal data of Data Principals with business entities.

Because AI training requires "privacy-preserving pooling" of aggregate data, the current consent-based framework is flagged as insufficient and as the cause of the "Data Deadlock". Having identified the privacy challenges in data sharing, the document goes on to recommend a new architecture.

The proposed Digital Public Infrastructure is meant to move away from bilateral agreements to a collaboration between data providers and AI model developers, supported by enabling participants such as cloud service providers, model providers and certifiers.

### The ecosystem roles

According to the document, DEPA for Training is made up of distinct but interdependent participants, each with a defined role:

1. **Data Principals**, who "may not interact directly with the training network". Their interests are to be protected through an institutional governance structure that operationalises electronic contracts, confidential computing and differential privacy.
2. **Data Collabs**, which bridge the gap between Data Principals and the AI innovation network. These intermediaries, which may be cooperatives, industry associations or research consortia, work with communities to aggregate, standardise and organise data from multiple sources. They "may be responsible for obtaining primary consent where applicable". *Naavi's note: this cuts into the role of the Consent Manager under DPDPA.*
3. **Training Data Providers**, the institutions that source, curate and prepare datasets for AI training. A TDP must apply privacy-preserving transformations and ensure that datasets meet technical and regulatory requirements for de-identification. TDPs include hospitals,

financial institutions and public agencies. They register with a central registry and are bound by its governance standards, enforced through electronic contracts.

4. **Training Data Consumers**, the innovators who access these datasets to develop, refine or evaluate AI models. TDCs must operate within the purpose limitation and ethical conduct defined in their data-sharing agreements, either with Data Collabs that federate datasets for a sector or with individual TDPs. *Naavi's note: these correspond to the research organisations contemplated under DPDPA.*

### Privacy and security by design

The document treats two pillars as the basis of privacy and security by design: rigorous application of **Differential Privacy** and comprehensive enforcement of a **Zero-Trust Security Framework**.

It argues that traditional anonymisation is not sufficient for training AI models, while overly anonymised data loses its usefulness for training. Differential Privacy is therefore recommended. To complement it, a Zero-Trust framework would strictly verify every data access request and protect data at all stages through encryption, with continuous auditing and real-time monitoring.

### Economics

The document claims that the data deadlock costs trillions of dollars in unrealised revenue and productivity. No specific evidence is offered for this figure.

### The proposed workflow

Ministries and other participants are expected to publish a Data Sharing Procedure, and a phased implementation roadmap is proposed. The framework describes a five-stage process:

1. **Data registration and encryption.** The TDP registers and catalogues its dataset, applies anonymisation and encryption, and attaches metadata on origin, sensitivity and access restrictions.
2. **Data request and contract validation.** The TDC submits a request for a specific use case. An electronic contract management system checks it against the provider's policies and the governance rules.
3. **Secure model training.** Data aggregation and computation take place inside the Secure Data Environment. Raw data is not exposed to the consumer. Differential Privacy and immutable audit logs are used.
4. **Model validation and bias detection.** The trained model undergoes independent checks for bias, fairness, security, robustness, data leakage and explainability. Only models that pass proceed to pilot deployment and scaling.
5. **Deployment and continuous monitoring.** The model may be deployed for commercial or public use, subject to continuing monitoring, fairness and security logs, and traceability to the datasets and contracts used in training.

### Part C: Issues the draft leaves open

The proposal is ambitious, but several matters need legal and operational clarification. The third column, new in this edition, shows where AIGSI Version 3.0 already offers an answer that DEPA could adopt. Clause references are to the AIGSI Version 3.0 discussion paper.

| Issue | Question requiring clarification | Where AIGSI v3 offers an answer |
|---|---|---|
| **Legal basis** | What legal basis permits the original collection, and the subsequent processing or sharing, of personal data for AI training? | Part A: every obligation traced to DPDPA or ITA 2000. AIGSI does not create a legal basis; it requires one to be recorded. |
| **Anonymisation** | Who decides that data has been sufficiently anonymised, and who is responsible if re-identification becomes possible? | B.6: whoever supplied the data answers for its use. P2: unknown risk is treated as significant risk. |
| **Data ownership** | What rights do data providers, Data Principals and Data Collabs have over contributed datasets? | Copyright control area (D.I); LC-14 IP indemnity. Ownership itself is a policy question AIGSI leaves to contract and law. |
| **Commercial rights** | Who owns the resulting model, and how are intellectual property and commercial benefits shared? | MO-9: whoever fine-tunes, quantises or distils a model is the Model Owner of the modified model. LC-14 IP indemnity. |
| **Accountability** | How are legal responsibilities divided among the provider, the collective, the platform, the processor and the AI developer? | Four roles (B.1), backward liability chain, B.6 rules where roles blur, B.7 failure allocation. |
| **Model leakage** | Who is responsible if a trained model reveals personal or confidential information despite the safeguards? | B.7 failure allocation; release controls and containment testing (D.I); F.8 post-incident review. |
| **Cross-border use** | What restrictions apply when a model, an infrastructure provider or a consumer operates outside India? | B.5: recourse stops at a supplier beyond effective jurisdiction; DS-44 residual loss cover; LC-12 exit plan. |
| **Oversight** | How does the proposed consortium relate to existing statutory regulators and authorities? | F.0: one audit, three opinions (CERT-In, DPDPA, IT Act); DS-38 reporting matrix keeps statutory duties separate. |

The document's reference to de-identifying personal data before sharing needs particular attention. Pseudonymisation and anonymisation are not interchangeable. Pseudonymised information remains personal data wherever an individual can still be identified using additional information.

The legal consequences must be examined under the applicable law, and cannot be assumed away by the technical architecture.

## Part D: Issues for Data Fiduciaries and Independent Data Auditors

When a Data Fiduciary adopts this framework, it creates a new challenge for the Independent Data Auditors who assess DPDPA compliance. The auditor must be satisfied that personal data exported by the Data Fiduciary to the DEPA-AIT network does not create harm or risk for the Data Principals concerned.

Two fundamental questions arise:

1. Does the ability to share data securely automatically create the legal authority to share and process it?
2. Does making Indian datasets available through this architecture strengthen India's AI sovereignty, or could it let foreign-controlled entities capture the strategic and economic value of Indian data? This is a question of ethics and responsible use of AI.

The concerns below follow from these questions. Where AIGSI Version 3.0 has a clause that helps, it is noted under the heading "AIGSI v3".

### 1. Data availability is not authority to use data

DEPA-AIT brings TDPs and TDCs into a common techno-legal ecosystem. But under DPDPA, processing of personal data must rest on a ground recognised under Section 4, read with Sections 5, 6 and 7 as applicable. Being a registered DEPA participant does not create an additional ground. DPDPA recognises no general ground called "AI innovation", and gives no blanket permission to use personal data for AI training because the training may eventually benefit society.

The Data Fiduciary must first decide whether the proposed processing is lawful. That means answering questions such as:

- Was the data collected for a purpose that covers the proposed AI training?
- Does the original notice adequately describe that purpose?
- Is consent required, and if so, has valid consent been obtained?
- Can a specific legitimate use under Section 7 be established?
- Is any statutory exemption actually available?
- Do other laws, contractual obligations or sectoral regulations impose restrictions?

A data exchange platform cannot answer these questions. DEPA can facilitate access to data. It cannot manufacture a lawful basis for processing it.

### 2. Confusing pseudonymisation with anonymisation

The document rightly recognises that removing direct identifiers such as names, addresses and identification numbers does not prevent re-identification, because datasets can be linked and individuals identified through combinations of innocuous attributes. It therefore proposes more sophisticated techniques, particularly Differential Privacy.

Yet the same document requires de-identification with "pseudonymized keys held by the originating agency". If the originating agency keeps keys that can link the data back to an individual, can the

shared information automatically be treated as non-personal? It cannot. Pseudonymisation is a security technique. It is not anonymisation in the legal sense.

The document admits that combining datasets raises the risk of identification, and still relies heavily on the data provider's own responsibility to anonymise before sharing. That is not enough. A Data Fiduciary cannot discharge its statutory duty merely by declaring its dataset anonymised. It needs a defensible assessment of identification risk, covering linkage with other datasets, the nature of the information, the technical capability of likely recipients and the consequences of re-identification.

If a standardised DEPA certification of anonymisation is introduced, it must not become an automatic legal safe harbour.

> **AIGSI v3.** Principle P2 treats unknown risk as significant risk. An anonymisation claim that has not been tested against linkage is therefore an unknown risk, to be handled as significant. Under B.6, whoever supplied the data answers for its use, so the TDP cannot pass the anonymisation question to the platform.

### 3. Differential Privacy is no substitute for lawful processing

The proposal gives great weight to Differential Privacy, which is a valuable safeguard. But three separate questions must be kept apart:

1. Is the original collection and processing of personal data lawful?
2. Is the transformation and use of the data sufficiently privacy-preserving?
3. Does the resulting model or output create risks for individuals?

Differential Privacy mainly addresses the second and third questions. It does not answer the first. And a statement that Differential Privacy has been used does not by itself prove that privacy risk is controlled. The implementation must be independently evaluated.

### 4. Purpose limitation: can data collected for one purpose serve another?

This is perhaps the most serious DPDPA issue for an individual Data Fiduciary.

Section 5 and the DPDP Rules 2025 require a notice describing the personal data and the purposes of processing. A broad purpose such as "improving services" cannot be read as authorising every future AI training activity. The Data Fiduciary must examine whether the new purpose is covered by the original notice and lawful basis, or whether a fresh basis and a new notice or consent are needed.

AI training is also rarely a single, clearly defined activity. One dataset may be used to train several models, fine-tune existing models, validate algorithms, test competing products and build commercial applications. A general permission to contribute data to a training platform could thus open the door to many downstream purposes never explained to the Data Principal.

DEPA must not become a means for a Data Fiduciary to turn a narrowly collected dataset into a reusable commercial resource without checking the legal basis for each processing activity.

### 5. Consent: who obtains it, for what, and from whom?

The document introduces Data Collabs as intermediaries that may work with communities to aggregate, standardise and organise data, and that "may be responsible for obtaining primary consent where applicable". This raises several questions:

- If a patient consents to a hospital for treatment, can a Data Collab later treat that consent as permission for AI training?
- If a Data Collab obtains consent, is it acting as the hospital's agent, as a Data Processor, or as an independent Data Fiduciary?
- Who gives the notice required under DPDPA?
- Who enables withdrawal of consent?
- What happens when consent is withdrawn after the data is already in a training dataset?

These go to the heart of the DPDPA consent architecture. Consent under the Act must be free, specific, informed, unconditional and unambiguous, given by a clear affirmative action, and specific to a purpose. Consent taken once at collection cannot become a permanent authorisation for an open-ended range of AI applications. A person may agree to her health data being used for research into a particular disease, but not to its use in an unrelated commercial profiling system.

Withdrawal matters equally. A framework that can take data into a training environment but cannot respond meaningfully to withdrawal tilts the balance from the individual towards the convenience of AI developers.

The research exemption under Section 17(2)(b) also needs care. It is subject to statutory conditions and prescribed standards, and cannot be read as a blanket exemption for all commercial AI training.

### 6. Erasure after AI training

DEPA envisages a lifecycle in which data is contributed, transformed, used for training, validated and finally deployed through a model. What happens when a Data Principal exercises the right to erasure under Section 12?

Deleting the source record may be possible, and so may deleting the copy on a Data Exchange Platform. But what about:

- the transformed training dataset
- intermediate training files
- checkpoints and model weights
- embeddings and derived representations
- validation datasets
- backup copies
- models already deployed to customers?

Can the model be shown not to retain or reveal the individual's data? Can the contribution be removed? Must the model be retrained, fine-tuned or put through machine unlearning? Who pays for it?

DPDPA prescribes no universal technical method for removing an individual's contribution from a trained model, and erasing the source record does not erase its influence on the model. DEPA must solve this before any data is contributed. A Data Fiduciary should not enter a data-sharing arrangement without understanding what withdrawal, erasure, retention and model reuse will mean.

> **AIGSI v3.** The wider Model Owner definition (MO-9) helps answer "who pays". A TDC that fine-tunes, quantises or distils a model on DEPA data becomes the Model Owner of the modified model, and carries the Model Owner duties for it. The AI Bill of Materials required where models are routed or embedded (B.6) gives the traceability from model back to dataset that an erasure request needs.

### 7. A Secure Data Environment is not a legal sanctuary

The document relies heavily on SDEs: hardware-based secure enclaves, encryption, access controls and immutable logs. These can greatly reduce the risk of unauthorised access and leakage. They do not make every activity inside them lawful.

An unauthorised purpose stays unauthorised inside a secure enclave. A model trained on personal data without a valid legal basis does not become lawful because the data never left the enclave. Nor can a secure environment cure inappropriate data selection, excessive collection, unauthorised secondary use, inaccurate data, discriminatory training data, improper retention or unlawful disclosure of model outputs.

The distinction is simple. Security controls protect the processing environment. Data protection law decides whether the processing should happen at all, and on what conditions. The framework should say so explicitly.

> **AIGSI v3.** This distinction is built into the AIGSI audit. Under F.0 one engagement gives three separate opinions: a CERT-In assurance opinion on the security of the AI system, a DPDPA opinion on personal data, and an IT Act opinion on the security of non-personal data and systems. A clean security opinion on the SDE cannot be read as a DPDPA opinion on the processing.

### 8. The Data Fiduciary cannot outsource accountability to DEPA

The ecosystem involves Data Principals, Data Collabs, TDPs, TDCs, Data Exchange Platforms, SDE providers, technology service providers and an oversight body. That is a long chain of responsibility, and DPDPA does not recognise most of these as statutory categories.

The Act knows only Data Fiduciaries, which determine the purpose and means of processing, and Data Processors, which process on a fiduciary's behalf. A participant does not become a Data Processor because the framework calls it a service provider. An AI developer that receives data and independently decides how to train a commercial model is unlikely to be merely a processor for the originating institution. Depending on the arrangement, it may be a Data Fiduciary in its own right. Likewise, a Data Collab that aggregates data from several institutions and decides how it will be used may carry responsibilities that no contractual label can settle.

Section 8(1) makes the Data Fiduciary responsible for processing done by it or on its behalf by a Data Processor, and Section 8(2) requires a valid contract to engage a processor. An electronic contract

saying that the platform, exchange or AI developer will take care of compliance cannot remove that statutory responsibility. A contract can allocate operational duties and provide indemnities. It cannot override the statute.

For each processing activity, the framework must therefore identify:

- who determines the purpose
- who determines the essential means
- who acts on behalf of another entity
- who is an independent Data Fiduciary
- who responds to Data Principal rights
- who handles breach notification
- who decides on erasure and retention

Without this allocation, the framework risks creating an accountability vacuum.

> **AIGSI v3.** AIGSI's rules for blurred roles (B.6) apply directly to DEPA: a party that fine-tunes a model becomes its Model Owner; an integrator that recommends a model or host answers for that choice; where one organisation holds several roles, an internal agreement names an owner and a Human Handler for each. B.7 then gives guidance on which party should answer for which kind of failure, for use in drafting indemnities. None of this changes the Data Fiduciary's statutory accountability; it fills in the chain behind it.

### 9. Data sovereignty: a gateway for foreign control of Indian data?

The document speaks of India's ambition to build its own AI ecosystem, yet the architecture is not reserved for Indian AI developers. TDCs may be start-ups, researchers or enterprises, and the framework openly envisages global collaboration, cross-border research and international data-sharing arrangements.

International collaboration is welcome in science and technology, and Indian institutions should take part in global research. But sharing data on terms that preserve Indian control is very different from an architecture in which foreign entities may capture the economic and strategic benefit of Indian data without matching Indian control.

The draft sets no Indian ownership or nationality requirement for TDCs. Nor does it provide any framework for keeping Indian control over the resulting models, derived intelligence and commercial value. This is a significant omission.

An institution created to serve a national data need can, over time, pass into the control of a global information company. The law may regulate its operations, access to information and security. But who controls the company, who owns the technology, who builds the analytics and who profits from the intelligence are separate questions. CIBIL showed that an Indian regulatory framework does not by itself guarantee Indian ownership or strategic control over an institution handling nationally important information. That experience should be treated as a warning while designing DEPA.

### Sovereignty is more than localisation

Data sovereignty cannot be reduced to whether the server sits in India. Localisation helps with some jurisdictional and security concerns, but it does not keep national control over the data ecosystem. There are at least four dimensions:

1. **Physical control:** where the original dataset is stored and processed.
2. **Access control:** who can access the dataset and authorise its use.
3. **Intellectual property control:** who owns the model, algorithms and derived IP created from the dataset.
4. **Economic control:** who receives the commercial benefit of the intelligence extracted.

A framework that protects only the first two may still let the strategic value of Indian data migrate to foreign-controlled entities. This matters most in healthcare, financial services, agriculture, education, geospatial information and public administration. Such datasets reveal India's population, consumption patterns, disease distribution, credit behaviour, agricultural productivity, geographical vulnerabilities and economic activity. Combined and analysed at scale, they become a strategic national resource.

### DPDPA alone does not secure sovereignty

DPDPA is a personal data protection law, not a national data sovereignty law. Section 16 allows the Central Government to restrict transfers to notified countries or territories, alongside other legal restrictions. It does not require personal data to stay in India.

A Data Fiduciary should not assume that DPDPA compliance keeps the economic and strategic value of Indian data under Indian control. Some datasets will fall outside DPDPA altogether, because they are genuinely anonymised or are not personal data, but their strategic importance remains. That is a policy gap. DEPA must therefore deal with personal data protection and with the wider question of national data governance.

### Where the liability chain ends

AIGSI Version 3.0 adds a point that sharpens the sovereignty concern. Its new section B.5 records where backward recourse actually stops. Two of its cases describe exactly what DEPA could produce:

- **A supplier beyond effective jurisdiction.** If a foreign TDC misuses data or releases a leaking model, an Indian contributor's contractual remedy may be unenforceable in practice. Recourse then stops, in practice, at the Indian party.
- **An open-weight model under an as-is licence.** If a model trained on Indian data is published with open weights, recourse stops at whoever chose to use it. Nobody upstream answers for it.

In both cases the Indian TDP, and the Indian deployer that later uses the model, are left holding the risk while the value has moved abroad. AIGSI's answer is that a choice made without recourse is a liability decision, to be recorded and approved at the level that accepts the risk, with residual loss covered by insurance or reserves (DS-44) and an exit plan in the contract (LC-12). DEPA should require the same of every TDP before data leaves its control.

### Safeguards to build in

1. **Indian control over critical national datasets.** Data relating to banking, healthcare, public administration, national infrastructure and geospatial intelligence should carry clearly defined access and control conditions, not left to individual providers.
2. **Eligibility and scrutiny of foreign participants.** Foreign entities need not be excluded, but foreign-controlled participants should face scrutiny of ownership, beneficial ownership, jurisdiction and security, especially for strategic datasets.
3. **Ownership and control of trained models.** Every contract should state who owns the resulting models, weights, derived datasets, improvements and IP, and should address commercial exploitation outside India. Under AIGSI MO-9, a TDC that fine-tunes or distils a model on Indian data becomes its Model Owner; the contract should say what that ownership carries with it.
4. **Restrictions on onward transfer and affiliate access.** A foreign participant should not pass data, derived information, model artefacts or access rights to its parent, subsidiaries, affiliates or third parties without authorisation.
5. **Indian jurisdiction and enforceability.** Contracts should provide for enforceable Indian jurisdiction, audit access, regulatory cooperation, dispute resolution and remedies, backed by practical enforcement capability. Where enforceability is doubtful, residual loss should be covered (AIGSI B.5, DS-44, LC-13).
6. **Governance of Data Exchange Platforms.** Ownership and control of critical platforms should be disclosed, with safeguards against concentration of control and provision for portability, continuity, exit and independent audit.
7. **Benefit-sharing.** Where Indian datasets contribute materially to commercially valuable models, Indian institutions, contributors and the domestic AI ecosystem should receive a fair share of the value.
8. **Exit and change-of-control provisions.** A change in ownership, control or jurisdiction of a platform or a major TDC should trigger review of existing permissions and contracts.
9. **No open-weight release without approval.** Release of model weights trained on DEPA data should need the TDP's express approval, recorded as a liability decision under AIGSI B.5.

These safeguards still allow international collaboration. They ensure India does not give up control over strategically important data assets in the process.

## 10. Society-wide aggregation and collective profiling

The societal implications of DEPA may be graver than the compliance risk to individual institutions. The proposal aims to combine data across sectors, which is exactly what makes it attractive for AI, and exactly what enables profiling of individuals and communities in ways never contemplated at collection.

Combine banking transactions, health records, education records, geospatial information, insurance data and consumer behaviour. Even without direct identifiers, the result can reveal income, health, financial distress, social behaviour, location and employment. AI systems may learn correlations that no single dataset shows. Such models may then decide credit eligibility, insurance pricing, employment suitability or access to public services.

The individual may never know that data from a hospital, bank or government department fed a model that later decided something about her. De-identified training data does not remove the need to examine what the deployed model does. A national AI data ecosystem must include safeguards against discriminatory profiling, group harms, inappropriate inference and misuse of model outputs.

> **AIGSI v3.** Version 3.0 adds fairness testing and protection of vulnerable persons as deployer control areas (D.I), with matching upstream duties for developers. A deployer using a DEPA-trained model to decide credit, insurance or access to services would have to test it for group harm, not rely on DEPA's Step 4 validation alone.

### 11. Electronic contracts cannot replace statutory compliance

Machine-readable electronic contracts, specifying permitted purposes, security requirements and access conditions and enforced by technology, are a welcome innovation. But a contract between a TDP and a TDC is not a contract with every Data Principal whose data sits in the dataset. The Data Principal is not bound by it, and it cannot extinguish her statutory rights.

An electronic contract can record that the AI developer will not use data for an unauthorised purpose. It cannot prove that the original Data Fiduciary was entitled to share the data in the first place. It can require deletion after a set period, but it cannot settle statutory retention obligations or override a Data Principal's rights.

The framework should keep four layers of governance distinct:

1. the legal basis for processing personal data
2. the contractual conditions governing access and use
3. the technical controls that enforce those conditions
4. the accountability of each statutory Data Fiduciary

They must work together. None can replace another.

> **AIGSI v3.** AIGSI holds the same line. Under LC-1 the affected person's remedy against the deployer never depends on what recourse exists upstream. Under LC-5 liability caps do not limit statutory liability, and caps, exclusions and carve-outs must be disclosed before contract. New section A.5 maps the liability doctrines to DPDPA, ITA 2000, the Indian Contract Act and the Consumer Protection Act, so that contract recourse and statutory liability are kept apart.

### 12. The oversight body: a parallel compliance regime?

The document proposes an oversight body to set standards, register participants, oversee data-sharing procedures and facilitate dispute resolution, aligned with the IT Act 2000 and DPDPA 2023. What is its legal status? A statutory regulator, a consortium, an industry self-regulatory organisation, a government-authorised platform operator, or some mix?

DPDPA vests statutory functions in the Data Protection Board of India. A DEPA oversight body cannot assume or displace those powers. There is also a real risk of confusion if a participant treats a DEPA certification as proof of DPDPA compliance. Organisations would then face two assessments,

one under DEPA and one under DPDPA, and a good result in the first could breed unjustified confidence about the second.

The governance arrangement should state plainly that registration, certification, security audits and participation in DEPA confer no exemption or immunity under DPDPA.

> **AIGSI v3.** AIGSI's principle is that organisations should not be audited twice for the same system. Under F.0 an AIGSI audit is conducted as a CERT-In AI security audit, with AIGSI's DPDPA-centric and ITA 2000-centric parts forming the external-requirements layer of CERT-In's Statement of Applicability. Rather than build its own certification, DEPA could accept such an audit, with its three separate opinions, as the evidence it needs from participants.

### 13. Additional obligations for Significant Data Fiduciaries

This matters most for organisations classified as Significant Data Fiduciaries (SDFs): banks, hospitals, insurers, large platforms and others whose scale or manner of processing draws enhanced scrutiny. Section 10 requires SDFs to appoint a Data Protection Officer and an independent data auditor, among other measures, and the DPDP Rules 2025 add obligations on risk assessment and audit.

If an SDF contributes personal data to a DEPA training project, the decision cannot be treated as routine technology outsourcing. The SDF must assess the whole processing chain:

- the legal basis for the contribution
- the classification and sensitivity of the data
- the risk of re-identification
- the identity and statutory role of each participant
- the security and privacy controls of the SDE
- the possibility of model leakage
- contractual restrictions on downstream use
- the ability to respond to Data Principal rights
- the consequences of a breach or misuse
- ownership and control of the resulting models and derived intelligence
- where its recourse ends, and how residual loss is covered

The independent data auditor should examine the legal and governance basis for the SDF's participation, and not merely the platform's security certificate. An SDF cannot excuse a failure of its own obligations by pointing to the DEPA platform or an external provider's certification.

> **AIGSI v3.** The AIGSI audit has branches for Significant Data Fiduciaries and for RBI-regulated entities. Version 3.0 adopts CERT-In's grading of findings and remediation timelines (F.6A) and its rules on auditor competence and independence, with added legal competence for the DPDPA and IT Act parts (F.6B). These give the SDF's independent data auditor a recognised way to grade what it finds in a DEPA contribution.

## 14. What a Data Fiduciary should do before joining DEPA

DEPA should not be rejected. But no Data Fiduciary should join simply because it is presented as a national architecture for responsible AI. Before contributing any dataset containing personal data, the organisation should complete a structured legal and techno-legal assessment. I recommend these minimum checks:

1. **Data provenance.** Identify the source of every dataset, the purpose of collection, the notice given, the legal basis and any restrictions attached.
2. **Personal data classification.** Determine whether the dataset contains personal data, pseudonymised data, genuinely anonymised information, or a mix.
3. **Legal basis.** Establish the specific basis under Sections 4, 6 or 7, as applicable. Do not assume AI training is covered by the original purpose.
4. **Role determination.** Decide whether the organisation acts as a Data Fiduciary, a Data Processor or another participant, based on actual functions rather than contractual labels. Map each DEPA participant to its AIGSI role (Part E).
5. **Re-identification risk.** Assess the chance of identification through linkage, and the effectiveness of the proposed anonymisation and Differential Privacy.
6. **Contracts.** Ensure agreements cover permitted purposes, prohibited uses, retention, onward-sharing limits, security obligations, audit rights, incident reporting and liability, including disclosed caps and exclusions, insurance and IP indemnity (AIGSI LC-5, LC-13, LC-14).
7. **Data Principal rights.** Establish how withdrawal of consent, access, correction, erasure and grievance redressal will work across the data-sharing and training lifecycle.
8. **Model risk.** Evaluate the risk of leakage, memorisation, inference, discriminatory outcomes and other harms from the trained model, and the release controls that apply before a model leaves the SDE.
9. **Security.** Evaluate the SDE's encryption, access controls, privileged access, logging, monitoring, vulnerability management and incident response.
10. **Sovereignty and value retention.** Determine who controls the platform, who owns the model and its improvements, whether foreign affiliates can reach derived information, and whether the Indian contributor keeps appropriate rights and benefits.
11. **Where recourse ends.** Record, for each downstream party, where contractual recourse stops (AIGSI B.5), and how residual loss will be covered (DS-44). Have the decision approved at the level that accepts the risk.
12. **Independent audit.** Obtain an independent assessment of legal, technical and organisational safeguards. For an SDF, integrate it with the statutory audit.

This assessment should be complete before the first dataset is transferred or made accessible.

## Part E: How AIGSI Version 3.0 can fill the gap

DEPA is concerned mainly with enabling data access and secure AI training. It is not a complete compliance framework for the Data Fiduciary, which is why an integrated approach to data governance and AI governance is needed.

### What AIGSI is, and the status of Version 3.0

AIGSI (Artificial Intelligence Governance Standard of India) is FDPPI's AI governance standard and an enhancement of DGPSI-AI. Version 2.0 (8 November 2026) unified AIGSI with the RBI Guidance on Model Risk Management, CERT-In's guidelines on AI-accelerated vulnerability protection and on AI security auditing, and the DSCI paper on AI liability. It anchored every obligation in DPDPA or ITA 2000, set liability to flow backwards from the deployer to the developer and up to the model owner through a contract at each step, and split the audit into a DPDPA-centric part and an ITA 2000-centric part with a common core.

Version 3.0 has been prepared as a discussion paper for the Special Interest Group on AI Governance that FDPPI is forming. It is a draft for discussion, and Version 2.0 remains the reference text until the SIG concludes. Version 3.0 is backward-compatible: every earlier clause keeps its number.

### The relationship between the instruments

DEPA-AIT and AIGSI work at different layers. DEPA is a data-access architecture. It decides how data moves from a Training Data Provider into a Secure Data Environment and how a model comes out. AIGSI is an organisational governance standard. It decides whether each participant has governed its own part of that journey, and who answers for it when something goes wrong.

| Instrument | Role |
|---|---|
| **DEPA** | Enables data sharing for AI training |
| **DPDPA** | Governs lawful processing of personal data |
| **ITA 2000** | Governs security of electronic records and systems, including non-personal data |
| **DGPSI-AI** | Supports organisational governance of AI-related processing |
| **AIGSI** | Addresses the wider governance of AI systems across the lifecycle, and the liability chain behind them |

Each has its own function, and none can stand in for another.

### Mapping DEPA participants to AIGSI roles

AIGSI recognises four roles: Deployer, Developer, Model Owner and Infrastructure Provider. Version 3.0 divides the Developer into three sub-roles (product companies, tool providers and integrators) and widens the Model Owner to include anyone who fine-tunes, quantises, distils or republishes a model. DEPA's participants map onto these roles as follows.

| DEPA participant | AIGSI role | What AIGSI Version 3.0 asks of it |
|---|---|---|
| **Training Data Provider (bank, hospital, public agency)** | Data supplier; Data Fiduciary under DPDPA | Lawful basis and purpose check before contribution; DPDPA-centric audit of outgoing data; answers for the use of the data it supplied (B.6); records where its recourse ends (B.5) |
| **Data Collab** | Integrator (Developer sub-role) or Data Fiduciary, by function | Answers for the choices it recommends (B.6, DV-39); clear statutory role; internal owner and Human Handler for each role it holds |
| **Data Exchange Platform and SDE provider** | Infrastructure Provider | ITA 2000-centric security audit against CERT-In's nine domains (F.7); security operations; upstream duties including IP-6 |
| **Training Data Consumer that trains or fine-tunes a model** | Developer, and Model Owner of the model it modifies (MO-9) | Model risk management; leakage and bias testing; release controls; Model Owner duties MO-1 to MO-9; disclosed caps, insurance and IP indemnity (LC-5, LC-13, LC-14) |
| **Foundation model provider whose model is fine-tuned** | Model Owner of the base model | End point of the liability chain; but B.5 records where recourse stops in practice (as-is licence, capped terms, jurisdiction) |
| **Bank, insurer or agency that deploys the trained model** | Deployer | Accountable to the affected person regardless of upstream recourse (LC-1); fairness testing, vulnerable persons, residual loss cover (D.I, DS-44); kill switch and Human Handler |

### Six gaps in DEPA-AIT that AIGSI Version 3.0 addresses

#### 1. The liability chain, and where it ends

DEPA has a chain of participants (TDP → Data Collab → DEP/SDE provider → TDC → model provider → deployer) but allocates no liability along it. AIGSI sets liability to flow backwards through a contract at each step. Version 3.0 adds what Version 2.0 did not say: that the chain often ends before it reaches the party at fault. B.5 lists the situations (as-is licences, capped standard terms, private deployment, insolvency, a supplier beyond jurisdiction). DS-44 and LC-13 require residual loss to be covered. For DEPA, where many TDCs will be foreign and many models may be released openly, this is the most important single addition.

#### 2. Who is who when roles blur

DEPA defines its participants by function in the network, not by legal role. AIGSI's B.6 supplies the rules DEPA lacks. A TDC that fine-tunes a model on DEPA data is the Model Owner of the modified model, and whoever supplied the data answers for its use. A Data Collab that recommends a model, host or architecture answers for that choice; if a client mandates a choice against its advice, its duty narrows to a written warning. A party that routes work across several models answers for evaluating

each of them and lists them in the AI Bill of Materials. B.7 then offers guidance on which party should answer for which kind of failure, for use in drafting indemnities.

### 3. Personal and non-personal data

AIGSI's audit has a DPDPA-centric part covering personal data and an ITA 2000-centric part covering the security of non-personal data and of the AI system itself. This matters because much of DEPA rests on the claim that data has been "anonymised". Data that leaves the scope of DPDPA does not leave the scope of governance: the ITA 2000-centric audit still covers it, together with the sovereignty concerns in point 9 of Part D.

### 4. Model validation and release

DEPA's Step 4 ("fit for use" validation) has no defined standard. AIGSI's model risk management requirements, drawn from the RBI Guidance, provide one: model inventory, tiering, validation, leakage and bias testing, and monitoring after deployment. Version 3.0 adds release controls and containment testing as control areas, and fairness testing for deployers. Together these answer the question of what a model must pass before it leaves the SDE and who signs off its release. New principle P17, "prevent first, halt last", sets the order: prevent harm through release controls; keep the independent kill switch (P9) as the last resort. P17 also warns that over-control is a risk of its own, which matters for a framework meant to enable innovation.

### 5. One audit, three opinions

DEPA requires SDE providers to be "registered" but sets no audit standard, and its proposed oversight body risks creating a second compliance regime. Version 3.0 aligns AIGSI with the CERT-In AI Security Auditing Guidelines. Under F.0 one engagement produces a CERT-In assurance opinion, a DPDPA opinion and an IT Act opinion. F.7 maps CERT-In's nine audit domains to AIGSI clauses, giving the auditor a benchmark for the SDE and the training pipeline. F.6A and F.6B adopt CERT-In's grading of findings, remediation timelines, and auditor competence and independence rules.

### 6. Copyright, synthetic content and incidents

DEPA is silent on the copyright status of training data and model outputs, and on what happens after an incident. Version 3.0 adds copyright and synthetic media as control areas, an IP indemnity clause (LC-14), and a post-incident review procedure (F.8) that turns DSCI's 21 interrogation lenses into questions mapped to AIGSI clauses. A model that leaks personal data after leaving the SDE would be reviewed through that procedure, and the reporting matrix (DS-38) keeps DPDPA, CERT-In and sectoral reporting as separate duties.

### Alignment with MeitY's own guidelines

Part G.2 of AIGSI Version 3.0 maps the standard to MeitY's India AI Governance Guidelines (2025), as well as to ISO/IEC 42001 and the NIST AI RMF. MeitY's DEPA handbook and MeitY's AI governance guidelines should point in the same direction. An organisation that follows AIGSI can show alignment with both.

### In short

A DEPA participant that also adopts AIGSI can show an auditor, a regulator or the Data Protection Board that it has done more than use a secure platform: it has governed its own role in the data-to-model lifecycle, it knows where its recourse ends, and it has covered the gap. DEPA should therefore recognise such organisational standards explicitly, and not treat its own registration or certification as a substitute for them.

Professionals who wish to help refine AIGSI Version 3.0, including its application to frameworks such as DEPA-AIT, are invited to join FDPPI's Special Interest Group on AI Governance.

## Part F: What MeitY should correct before finalising the framework

The final version should carry explicit safeguards for both statutory compliance and national data sovereignty. I recommend that it state this principle:

> Participation in DEPA shall not, by itself, constitute a lawful basis for processing personal data under the DPDPA, nor shall registration, certification or use of a Secure Data Environment relieve any Data Fiduciary of its statutory obligations.

Every Training Data Provider should also be required to document:

1. whether the dataset contains personal data
2. the legal basis for the proposed processing
3. the purpose for which the data is being made available
4. the assessment supporting any claim of anonymisation
5. the identity and legal role of each participant
6. the safeguards for responding to Data Principal rights
7. responsibility for breach notification and remediation
8. arrangements for retention, deletion and model-level risk management
9. where contractual recourse against each downstream party ends, and how residual loss is covered

In addition, the architecture should include a separate data sovereignty policy covering:

- ownership and beneficial ownership of critical platform operators
- eligibility and scrutiny of foreign-controlled participants
- control over access to strategically important Indian datasets
- ownership and commercial use of trained models and derived intelligence
- restrictions on onward transfer to foreign affiliates
- Indian jurisdiction and enforceability
- change-of-control and exit arrangements
- benefit-sharing with Indian data contributors and the domestic AI ecosystem
- approval by the data provider before weights of a model trained on its data are released openly

Drawing on AIGSI Version 3.0, the framework should also:

- **define liability along the chain**, assigning each DEPA participant a role (Deployer, Developer, Model Owner or Infrastructure Provider) and requiring contracts that disclose caps and exclusions and provide for insurance and IP indemnity
- **treat a TDC that fine-tunes or distils a model as its Model Owner**, with the duties that go with it
- **set release controls** that a model must pass before it leaves the SDE, with a named person who signs off the release

- **accept one independent audit** producing separate security, DPDPA and IT Act opinions, instead of creating a parallel DEPA certification
- **require fairness testing and protection of vulnerable persons** before a DEPA-trained model is used to make decisions about individuals
- **recognise organisational governance standards** such as AIGSI explicitly

International collaboration should remain possible, on transparent conditions that protect individuals, Indian institutions and the national interest.

### Conclusion: India must not become a mere supplier of training data

India needs a national architecture for AI training, and nobody disputes that. Indian AI models built on Indian datasets can bring enormous economic and social value to healthcare, banking, education, agriculture and public administration.

But the drive to make India data-rich for AI must not end with individuals losing control over the lawful use of their personal data, or Indian institutions losing control over the value generated from the data they contribute. DPDPA was enacted to balance both: the individual's right to protect personal data, and the need to process it for lawful purposes.

DEPA has made a useful contribution by identifying technical mechanisms for sharing and processing data with reduced exposure. Its success should be judged by more than the volume of data released, the number of models trained or the value generated. The test must also include whether individual rights are protected, whether each Data Fiduciary can demonstrate compliance, and whether India retains real control over the strategic value of its data.

CIBIL reminds us that an Indian institution holding nationally important information can pass into foreign corporate control. The AI era multiplies the scope for value extraction, because the benefit now lies in the underlying data and also in the models, algorithms and intelligence built from it. AIGSI Version 3.0 adds a further warning: when something goes wrong, the chain of recourse may stop at the Indian border, leaving Indian institutions and citizens to bear the loss.

We face two dangers. One is that the convenience of a national data-sharing infrastructure becomes an excuse to weaken purpose limitation, consent, accountability and individual rights. The other, equally serious, is that India supplies the raw material for global AI while ownership and profits accumulate elsewhere.

A Secure Data Environment may stop a hacker from stealing a database. Differential Privacy may reduce the chance of identifying an individual from model outputs. An electronic contract may restrain a recipient from misusing data. None of them answers the fundamental legal question:

***Was the Data Fiduciary entitled to use this individual's personal data for this purpose in the first place?***

And none answers the equally important national question:

***Who will own, control and benefit from the intelligence created from Indian data, and who will answer for it when it causes harm?***

These questions must sit at the centre of India's AI data governance architecture. Otherwise India may supply the data, foreign entities may own the intelligence, and Indian society may end up paying for access to AI capabilities built from its own data.

Data empowerment without data accountability is not data protection. Data sharing without data sovereignty may become data extraction.

**Naavi**

## Annex: Questions for Discussion

Professionals reviewing this paper are invited to answer the questions below. Please quote the question number, say whether you agree with the position taken in the paper, and give your reasons. Partial responses are welcome. Responses may be sent to naavi@fdppi.in with the subject line "DEPA for AI Training".

Questions 1 to 35 are unchanged from the first edition. Questions 36 to 41 are new, and Question 29 has been updated to refer to AIGSI Version 3.0.

### A. Policy and purpose

1. Is the "Data Deadlock" a genuine problem that needs a new national architecture, or can it be solved within the existing consent framework of DPDPA?
2. Should access to DEPA-AIT be limited to Indian-owned and Indian-controlled entities? If foreign participants are allowed, on what conditions?
3. Should Government-held datasets (for example Nikshay, e-Sanjeevani, GST filings, Academic Bank of Credits, genomic data) be made available through DEPA-AIT at all? Which categories should be excluded outright?
4. Should datasets created with public money carry a royalty or benefit-sharing obligation? How should their value be determined, in line with the CAG's observations on valuing data?
5. Should DEPA-AIT remain a voluntary handbook, or be given statutory backing through rules under DPDPA or the IT Act 2000?

### B. Legal basis under DPDPA

6. What lawful basis can a Training Data Provider rely on to contribute personal data for AI training: consent, a legitimate use under Section 7, or the research exemption under Section 17(2)(b)?
7. Can the Section 17(2)(b) exemption cover commercial AI training? What conditions should the Rules prescribe for it?
8. Is pseudonymised data, where the originating agency keeps the keys, still personal data? Who should certify that a dataset is anonymised, and against what standard?
9. Can a model trained on personal data, even with Differential Privacy, itself contain personal data? Should model weights be treated as personal data for any purpose?
10. How should the rights to withdraw consent and to erasure (Sections 6(4) and 12) apply once data has been used to train a model?
11. Can the education use cases (adaptive learning, skills passports) be reconciled with the restrictions on children's data under Section 9?

### C. Roles and accountability

12. How should each DEPA participant (Training Data Provider, Data Collab, Data Exchange Platform, Secure Data Environment provider, Training Data Consumer, model provider) be classified under DPDPA: Data Fiduciary, Data Processor or otherwise?

13. If a Data Collab obtains consent on behalf of Data Principals, is it acting as a Consent Manager? Should it then register under Rule 4?
14. Who should be liable if a trained model leaks personal data or enables re-identification? Can an electronic contract shift that liability?
15. Should Data Collabs and Data Exchange Platforms be notified as Significant Data Fiduciaries, with DPIA and independent data audit obligations?
16. What should be the legal status of the proposed oversight body, and how should it relate to the Data Protection Board of India?

### D. Technical safeguards

17. Should downloading datasets (Option 1 of the handbook) be permitted for any data derived from personal data, or should all training be confined to Secure Data Environments?
18. Who should set the Differential Privacy budget (epsilon) for each sector, and should the values be published?
19. What tests should a model pass before it leaves the Secure Data Environment (for example extraction and membership-inference tests), and who should sign off on its release?
20. Should Secure Data Environment infrastructure be Indian-controlled, with hardware attestation verifiable in India?

### E. Data sovereignty

21. Who should own the model weights and intellectual property of models trained on Indian datasets?
22. Should foreign AI developers be allowed to fine-tune their foundation models on Indian datasets and deploy them globally? If so, on what terms?
23. Should onward transfer of data, derived information or model artefacts to foreign parent companies and affiliates be restricted?
24. Should a change in ownership or control of a platform or a major Training Data Consumer automatically trigger a review of its access, given the CIBIL experience?
25. Who should negotiate the cross-border data agreements envisaged in the roadmap: the Central Government alone, or platform operators as well?
26. Is a model of accredited “Indian Data Training Companies”, acting as a protective layer between data providers and AI developers, a workable alternative?

### F. Data Fiduciaries and Independent Data Auditors

27. What should an Independent Data Auditor verify before accepting a Data Fiduciary’s decision to contribute data to DEPA-AIT?
28. Should DEPA registration or certification be accepted as evidence of “reasonable security safeguards” under Section 8(5)?
29. How should standards such as DGPSI-AI and AIGSI Version 3.0 be applied to DEPA participants? Should DEPA recognise such organisational standards formally? *(updated)*

30. What additional checks should a Significant Data Fiduciary, such as a bank or hospital, complete before joining DEPA-AIT?

### G. Society and the Data Principal

31. What safeguards are needed against collective profiling and group harms when data from banking, health, education and geospatial sources is combined?
32. Should Data Principals have a public registry of datasets and consumers, a collective opt-out, and a grievance channel within DEPA?
33. Does DEPA-AIT, as drafted, risk being seen as a dilution of DPDPA, and could it affect the Government's defence of the Act before the Supreme Court?

### H. Overall view

34. On balance, should MeitY adopt DEPA-AIT as drafted, adopt it with modifications, or withdraw it for a fresh consultation?
35. What are the three most important changes you would recommend to MeitY?

### I. AIGSI Version 3.0 and DEPA-AIT (new)

36. Is the mapping of DEPA participants to AIGSI's four roles (Deployer, Developer, Model Owner, Infrastructure Provider) in Part E correct? Which participant is hardest to classify?
37. Should a Training Data Consumer that fine-tunes or distils a foundation model on Indian data be treated as the Model Owner of the modified model (AIGSI MO-9), with the duties that go with it?
38. Where a Training Data Consumer is beyond effective Indian jurisdiction, should the Training Data Provider be required to cover residual loss through insurance or reserves (AIGSI B.5, DS-44)? Who should bear that cost?
39. Should release of open weights of a model trained on DEPA data require the express approval of every Training Data Provider whose data was used?
40. Should DEPA accept a single independent audit producing separate CERT-In, DPDPA and IT Act opinions (AIGSI F.0) in place of its own certification of participants?
41. Would you be willing to take part in FDPPI's Special Interest Group on AI Governance, and which workstream interests you: liability and contracts, audit and assurance, technical controls, or legal and sectoral?

**Please complete this questionnaire online at here**




# Data Empowerment and Protection Architecture (DEPA) for AI Training

**A National Framework for**
**Data Empowered Ecosystems for shaping the Age of AI**

# Acknowledgements

At the Ministry of Electronics and Information Technology (MeitY), we are glad to have steered this important initiative. I would like to personally acknowledge the contributions of my colleagues, particularly the leadership shown by Shri Nand Kumaram, Joint Secretary, MeitY and P&CEO, NeGD and the dedicated efforts of Dr. Vageesh Tiwari, Director and Santosh Kr. Patel, Scientist-E (Digital Governance Division), MeitY, Shri Rajnish Kumar, COO, NeGD, Shri Bramhanand Jha, VP, NeGD and Officials of AI&ET Division, MeitY. Their support and collaboration were instrumental in this process.

The National E-Governance Division (NeGD) has made a commendable contribution in shaping and articulating the critical dimensions of data governance that inter-alia includes data sharing procedures crucial for AI development, and Data Exchange Platforms (including secure data environment) as safe and efficient mechanism for inter-agency data sharing. Its thoughtful approach has added practical artefacts of India's inclusive, resilient and vibrant digital governance ecosystem.

We would like to acknowledge and thank the contribution of the iSPIRT Foundation for their pioneering initiative and the immense effort invested in preparing this document. Special recognition is warranted for the visionary guidance of Sharad Sharma and the tireless contributions of Anshul Khandelwal, Harshit Kacholiya, Saumya Jetley, Shyam Sundaram, Sunu Engineer and Vikram Garg. This work was brought to fruition by the dedication of the entire iSPIRT volunteer team that worked on this vital project.

This document on data management procedure for AI training is intended to be a living resource that evolves alongside the DEPA framework. It aims to foster a vibrant discourse, and we welcome and invite actionable recommendations from individuals and institutions who are passionate about refining and co-creating this digital public infrastructure for the benefit of all.

S Krishnan

Secretary, Ministry of Electronics and IT

Government of India

अश्विनी वैष्णव
Ashwini Vaishnaw

सत्यमेव जयते

रेल, सूचना और प्रसारण एवं इलेक्ट्रॉनिकी
और सूचना प्रौद्योगिकी मंत्री
भारत सरकार
Minister of Railways, Information & Broadcasting
and Electronics & Information Technology
Government of India



**Foreward**

India is moving fast on digital transformation. This journey is guided by the vision of Hon'ble Prime Minister Shri Narendra Modi i.e democratisation of technology. Digital India Programme is helping take this vision to population scale.

India's Digital Public Infrastructure i.e India Stack has emerged as a framework with strong foundation. From JAM Trinity to UPI and ONDC, India Stack has set global benchmarks for using technology for public good.

Now, we are entering the age of AI. It can help solve complex national challenges and accelerate our journey towards a *Viksit Bharat*.

Data is the fuel for this AI revolution. The challenge is to realise the true potential of data while upholding the sanctity of a citizen's privacy as guaranteed by the constitution and Digital Personal Data Protection Act, 2023.

Data Empowerment and Protection Architecture (DEPA) for AI Training is our solution to the challenge. It provides for a consent-based framework and ensures a secure and user-controlled data sharing. It aims to helps startups, academia, & industry build world-class AI models tailored for India, by Indians.

DEPA supports the vision of "Make AI in India" and "Make AI work for India." It will drive innovation and create high-value jobs. It will help power our journey towards a trillion-dollar digital economy.

अश्विनी

(Ashwini Vaishnaw)

Electronics Niketan, 6, C.G.O. Complex, Lodhi Road, New Delhi-110003
Phone: 011-24369191, 24362626

जितिन प्रसाद
**JITIN PRASADA**

सत्यमेव जयते

राज्य मंत्री
वाणिज्य एवं उद्योग,
इलेक्ट्रॉनिकी और सूचना प्रौद्योगिकी
भारत सरकार

**Minister of State**
**Commerce & Industry,**
**Electronics and Information Technology**
**Government of India**



## MESSAGE

*India's journey in building its world-leading Digital Public Infrastructure (DPI) stack, from Aadhaar, UPI, Account Aggregator to DigiLocker, has set the global standard for inclusive digital empowerment. This has laid the foundation for our next great national endeavour: leading the global Artificial Intelligence revolution.*

*Our government's vision is clear: 'Make AI in India, and Make AI work for India.' To realize this vision, we are nurturing a robust AI ecosystem through the comprehensive IndiaAI Mission. A critical pillar of this mission is access to high- quality, large-scale data to train our models. However, this ambition carries a profound responsibility. The key challenge is to unlock the immense value held within siloed datasets while ensuring uncompromising adherence to the principles of privacy and citizen protection, as enshrined in our landmark Digital Personal Data Protection (DPDP) Act, 2023.*

*The Data Empowerment and Protection Architecture (DEPA) for Training represents our strategic, techno-legal response to this challenge. DEPA is both a technical architecture and an economic catalyst. By democratizing access to data, DEPA empowers our dynamic startup ecosystem, accelerates cutting-edge research in our academic institutions, and enables businesses, both large and small, to build sophisticated, India-specific AI solutions. It is a critical enabler for our goal of a $1 Trillion digital economy, strengthening the backbone of India's commerce and industry.*

*This framework is a testament to India's capability to build sophisticated, scalable, and trusted digital systems. It is a cornerstone of a 'Viksit Bharat' and solidifies our position as a responsible global leader in the age of Artificial Intelligence.*

**(Jitin Prasada)**

Digital India
Power To Empower

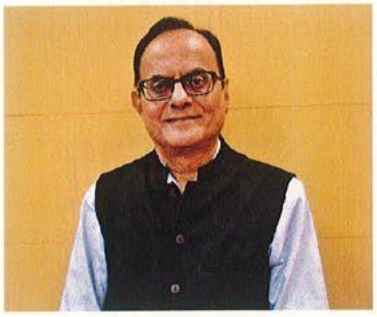
स्वच्छ भारत
एक कदम स्वच्छता की ओर

कार्यालय : इलेक्ट्रॉनिक्स निकेतन, 6, सी.जी.ओ. कॉम्प्लेक्स, नई दिल्ली–110003, दूरभाष : 011-24368757-58
**Office : Electronics Niketan , 6, C.G.O. Complex, New Delhi-110003, Tel. : 011-24368757-58**
**E-Mail : mos-eit@gov.in, Website : www.meity.gov.in**

अजय के. सूद
भारत सरकार के प्रमुख वैज्ञानिक सलाहकार
**Ajay K. Sood**
Principal Scientific Adviser to the Govt. of India

सत्यमेव जयते

विज्ञान भवन एनेक्सी
मौलाना आजाद मार्ग, नई दिल्ली–110011
Vigyan Bhawan Annexe
Maulana Azad Road, New Delhi - 110011
Tel. : +91-11-23022112
Fax: +91-11-23022113
E-mail : sood.ajay@gov.in
office-psa@nic.in
Website : www.psa.gov.in



## FOREWORD

India stands at the threshold of a new era, one defined by Artificial Intelligence and the data that fuels it. Our nation's journey in building world-class Digital Public Infrastructure-from Aadhaar and UPI to DigiLocker and the foundational Data Empowerment and Protection Architecture (DEPA) for inference (Account Aggregator)-has empowered millions and laid the groundwork for a trillion-dollar digital economy.

However, the next leap in value creation, particularly realizing the project $500 billion contribution of AI to our GDP, requires unlocking the immense potential currently trapped within siloed datasets. The challenge has been to enable access to aggregate data for AI model training without compromising the principles of privacy and security enshrined in the Indian legislation framework like the Digital Personal Data Protection (DPDP) Act, 2023.

DEPA training, the Digital Public Infrastructure for AI presented in this document offers an elegant, India-specific solution. By introducing Confidential Clean Rooms, Differential Privacy, Electronic Contracts and a collaborative governance model led by Self-Regulatory Organizations, it provides the necessary techno-legal rails for responsible data sharing.

This framework is more than just a technical architecture; it is a critical enabler for our startups, established industries, and research institutions. It democratizes access to data, allowing innovators across the spectrum to build sophisticated AI solutions tailored to India's unique needs.

DEPA reinforces India's position as a global leader in building digital infrastructure for citizen empowerment and economic growth, demonstrating a path that balances progress with protection.

Ajay Kumar Sood

**(Ajay K. Sood)**

**Dated: 27th August, 2025**

एस. कृष्णन, आई.ए.एस.
सचिव
S. Krishnan, I.A.S.
Secretary

सत्यमेव जयते

इलेक्ट्रॉनिकी और सूचना प्रौद्योगिकी मंत्रालय
भारत सरकार
Ministry of Electronics &
Information Technology (MeitY)
Government of India

# FOREWORD

The Ministry of Electronics and Information Technology (MeitY) has been at the forefront of architecting India's Digital Public Infrastructure, creating the foundational rails that power our vibrant digital economy. Our journey has been one of continuous innovation, moving from establishing digital identity and payments to building new, consent-based frameworks for data sharing. The Data Empowerment and Protection Architecture (DEPA) is a cornerstone of this next-generation architecture.

Technology must serve as a bridge between law and practice, and this is precisely the role that DEPA for Training is designed to fulfill. It provides a concrete, techno-legal pathway for organisations to collaborate on AI model training in a manner that is secure, auditable, and fully compliant with the principles of the DPDP Act.

The framework presented in this document is the result of extensive collaboration between government, industry, and academia. Its design, incorporating Confidential Clean Rooms and Differential Privacy, addresses the core challenge of enabling data analysis without data transfer, thereby minimising privacy risks. Furthermore, its federated governance model, led by Self Regulatory Organizations, ensures that the framework can adapt to the specific needs of different sectors while maintaining high standards of security and accountability.

This training document is an essential resource for all stakeholders in the digital ecosystem. It provides the necessary knowledge and tools to adopt this infrastructure, fostering a culture of responsible data stewardship. At MeitY, we are committed to building a trusted and innovative digital ecosystem, and we believe DEPA is a critical enabler of that vision, providing the certainty and clarity required for India to become a global leader in Artificial Intelligence.

(S. Krishnan)

Place: New Delhi
Date: 22nd October, 2025

Digital India
Power To Empower

इलेक्ट्रॉनिक्स निकेतन, 6, सी.जी.ओ. कॉम्पलेक्स, नई दिल्ली–110003 / Electronics Niketan, 6, C.G.O. Complex, New Delhi-110003
Tel. : 011-24364041 • email : secretary@meity.gov.in

# Foreword

## Shri Abhay Karandikar

**Secretary, Department of Science and Technology, Government of India**

For centuries, progress has been driven by theory and experimentation. The late twentieth century added large-scale simulation as a third pillar. Today, we stand firmly in the fourth paradigm of science: data-intensive discovery. In this new era, the ability to securely and ethically analyse vast, diverse datasets is not merely an advantage—it is the primary engine of innovation.

Historically, this meant building laboratories, observatories, and experimental testbeds. Today, it also means architecting a robust, trusted, and accessible national data infrastructure. For too long, Indian researchers have faced a critical "data chasm"—possessing world-class talent but often lacking access to the large-scale, high-quality, India-specific datasets required to solve our most pressing challenges.

The DEPA for Training framework is a direct and powerful response to this chasm. We see it not as a regulatory document, but as a fundamental piece of our National Infrastructure. By providing a techno-legal mechanism for privacy-preserving data collaboration, DEPA offers our scientists and innovators trusted access to the fuel needed for a new generation of discovery.

I call upon our vibrant community of researchers, academic institutions, and deep-tech startups to see this framework as an invitation. It is an invitation to ask bolder questions, build more ambitious models, and develop indigenous solutions for India's unique challenges in areas from precision agriculture and climate resilience to drug discovery and public health.

18th October 2025

# Foreword

## Shri Sibi George

**Secretary (West), Ministry of External Affairs, Government of India**

The world today stands on the threshold of a defining transformation, one that la powered by Artificial Intelligence (AI). This emerging force is not only reshaping global development, governance and security paradigms, but, and as our Hon'ble Prime Minister says, is writing the code for humanity. Al is projected to add over $15 trillion to the global economy by 2030, and it is often said that data is its new oil. India, with one of its largest and most diverse digital footprints in the world, generates volumes of data. Yet, the power of this data cannot be unlocked without ensuring trust, privacy and equity. The challenge before the international community therefore is to enable collaboration without compromising sovereignty and harness data responsibly without compromising citizens' rights and privacy.

It is with this vision that India's brainchild - the Data Empowerment and Protection Architecture (DEPA) was created. Rooted in our proven experience with Digital Public Infrastructure (DPI)-from Aadhaar and UP| to the Account Aggregator system, DEPA represents a new generation of digital frameworks that embed privacy, accountability and transparency into their very design. It is in line with our belief that the answers to complex global challenges lie not in imposing a singular world view, but in creating common platforms that enable diverse actors to collaborate with trust and confidence.

Where other global regulatory frameworks often swing between excessive controls or unchecked autonomies, India's model focuses on a citizen-centric, consent- based and techno-legal balancing architecture. It builds trust by law, by design and by technology, thereby epitomizing the principles of our constitutional democracy, even in the digital realm. It presents a replicable model for secure, decentralized and ethical data collaboration.

The "DEPA for Al Training Handbook" prepared meticulously by the Ministry of Electronics and Information Technology (MeitY) proves to be more than just a technical guide. It is a statement of intent and a positive contribution to the global thinking on data and Al governance. It reflects India's commitment to building a trusted, equitable and human-centric digital future. As the world steps forward to shape the discourse on the responsible use of data and Al, we hope that India's DEPA stands as a tangible symbol of our actions and our belief, that technology, when governed wisely, can unite rather than divide our shared digital planet.

27th October 2025

# Executive Summary

## Index

01 The Data Deadlock

02 DEPA for AI Training

03 A Techno-legal approach to AI Governance

04 Unleashing Sectoral Transformation

05 A Global Vision for Responsible innovation

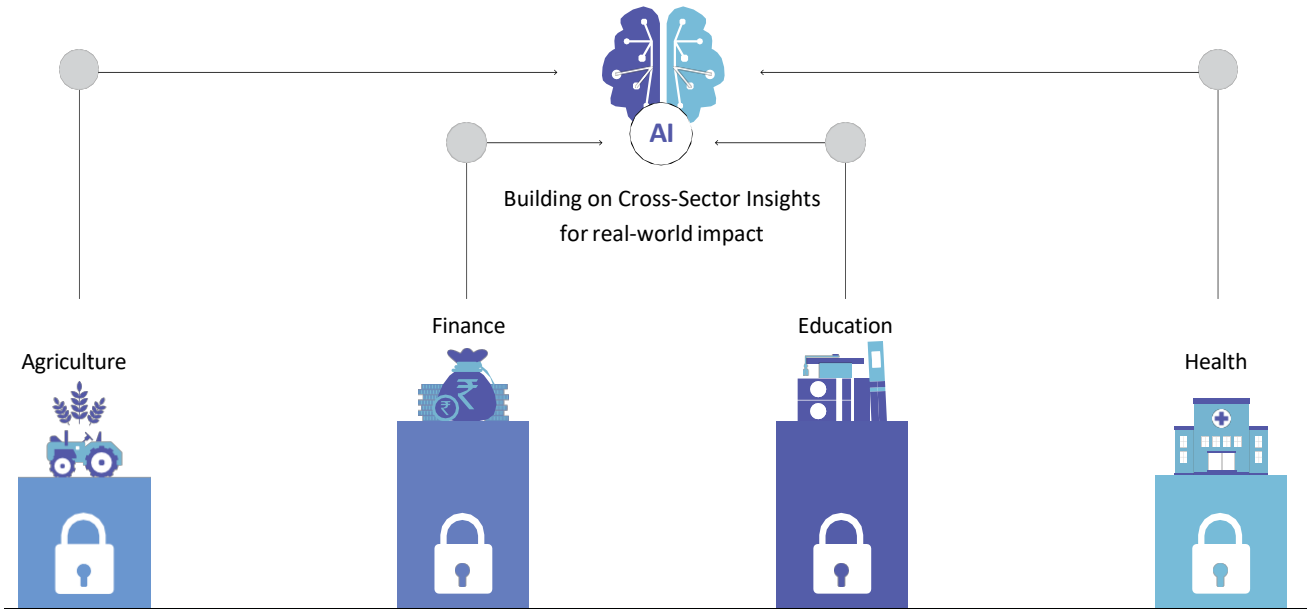
# 01 The Data Deadlock

Artificial Intelligence holds immense potential, but its power is fundamentally fueled by vast, diverse, and high-quality training data. AI needs to draw on collective patterns learned from aggregate data - spanning populations, domains, and time periods - to adapt to and solve for real world use cases. The scale and diversity of this data directly determine an AI model's capabilities, fairness, and reliability. India stands at a critical juncture, generating unprecedented volumes of digital data across sectors like healthcare, finance, and agriculture due to its world-class technology infrastructure. Notably, India is data rich. The critical challenge is to convert this wealth into real world solutions. The key to solving uniquely Indian challenges is to democratise data access for the building of specialised AI models.

Impactful, real-world AI models require complex, connected, real-world data to learn from. Data required to solve for the next generation of problems sits divided across domains, safeguarded behind different institutional and regulatory walls. We call this the problem of 'Data Deadlock,' a stalemate that stems from significant barriers observed across the globe. These include, among others, legal and regulatory uncertainty around data sharing given as the AI landscape is fast evolving, technical incompatibilities between systems that are both legacy and modern, and lack of institutional understanding on best practices for data security and privacy especially when traditional techniques of anonymisation and model training are increasingly insufficient in maintaining guarantees of privacy and confidentiality.

Data Deadlock is a critical challenge in the era of AI solutions. The consequences of inaction could be severe: stifled innovation, particularly for Indian startups and researchers; amplified societal biases embedded in AI models owing to their training on limited, foreign datasets; increased market concentration favoring large entities; and a failure to harness a key national strategic advantage. To overcome this deadlock and ensure steady progress of the data economy with benefits to the society at large, there is an urgent need for a data framework that can unlock data sharing and collaboration in a responsible manner, ensuring privacy, transparency, and trust.

**Unlocking India's AI potential requires intra and inter-sectoral collaboration on training data**



# 02 DEPA for AI Training

India's path to AI leadership demands a radical reimagining of our data infrastructure. DEPA Training is our proposed solution: India's DPI for AI helps break existing data collaboration barriers while ensuring privacy preserving mechanisms embedded into system design. Built on the ADEPTS framework — Accountability, Democratization, Evolvability, Privacy, Transparency, and Security — this architecture establishes the critical foundation of trust for India's data ecosystem.

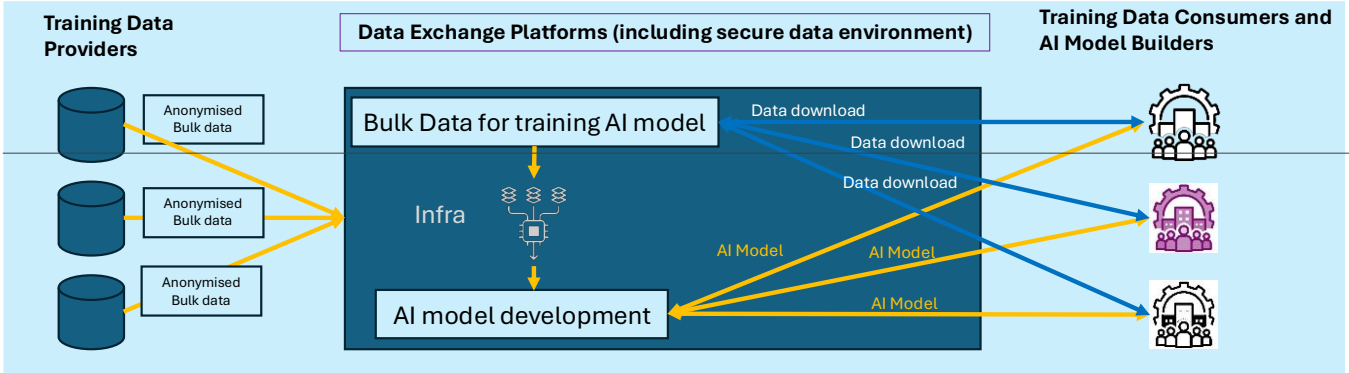
DEPA Training facilitates collaboration among key participants in the AI training lifecycle. Training Data Providers (TDPs) — institutions such as hospitals, banks, and public agencies holding valuable training data — collaborate with Training Data Consumers (TDCs) - encompassing innovators such as start-ups, researchers and enterprises, to bring innovative data-driven solutions for societal impact. Through this framework, TDPs and TDCs are brought together onto a common techno-legal platform where secure, private and compliant data exchange and AI modelling can proceed to foster new-age innovation.

This secure interaction relies on the core technical pillars: Data Exchange Platform, Electronic Contracts, Secure Data Environment (SDE), and Differential Privacy (DP). The secure data environment functions as neutral and secure environment for computations to happen in. Differential Privacy offers critical privacy guarantees during the data analysis and modelling phase, while Electronic Contracts provide a transparent, and automated framework for enforcing data sharing agreements and usage policies between participants.

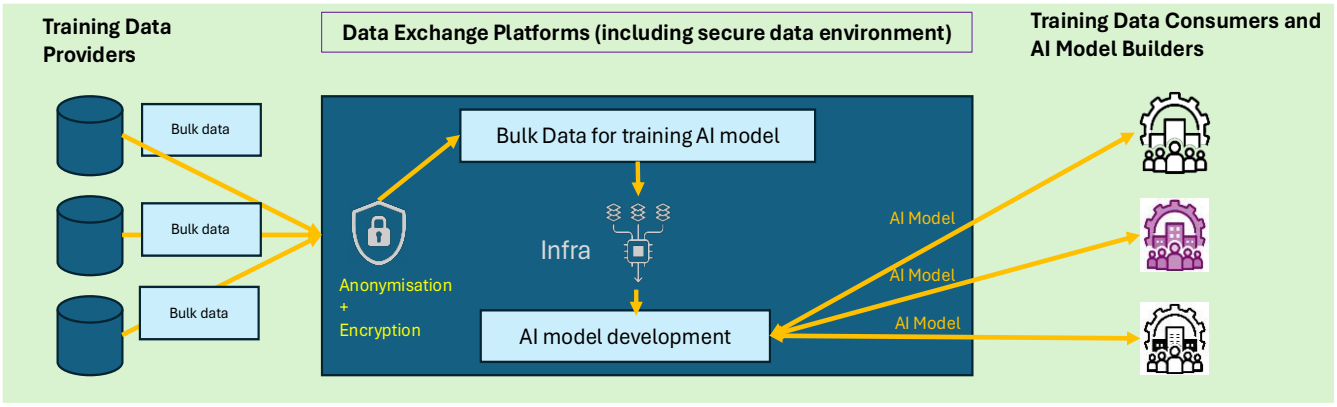
This framework provisions the use of Secure Data Environments along with Data Exchange Platforms for AI training, ensuring that TDPs' raw data remains protected in secure data environments and are unexposed to TDCs. Anonymisation of data is the responsibility entrusted to Training Data Providers. Training leveraging Differential Privacy can then occur within the secure data environment, allowing AI models to learn aggregate patterns while making it practically impossible to re-identify any individual data point, effectively addressing the shortcomings of traditional anonymization approaches. Governed by technologically enforceable contracts containing machine-readable attributes for data protection, security and use, and overseen by an appropriate regulatory organisation, this techno-legal architecture presents a robust pathway to safely unlock data for the advancement of India's AI ecosystem.

There can be two options. In the first option, the anonymisation is the responsibility of training data providers, while data consumers / data requestors get the data to download as well as the option to make use of secure data environment (of data exchange platform) for AI model training and testing. In the second option, the anonymisation and encryption of raw data (as provided by data providers) is handled by Data Exchange Platform, wherein, data consumers / data requestors get the access to data in secure data environment for AI model training and testing.

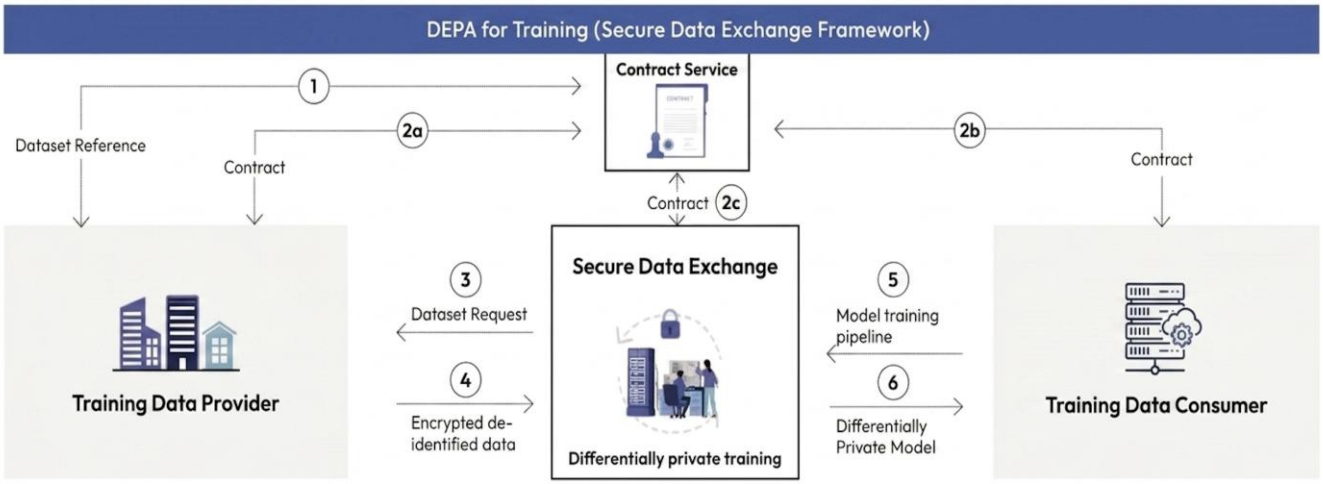
**Option 1: Data Exchange Platform getting anonymized data from data providers**




**Option 2: Secure Data Environment of Data Exchange Platform handling anonymisation of raw data**




Data Exchange Platform can play a role of data repository, where it allows data requestors to simply register and download the datasets. The need for registration can be decided by Data Providers based on the data classification (open, registered and restricted, as per NDSAP Policy, 2012). Data Exchange Platform (including secure data environment) can additionally provide the anonymisation and encryption / decryption as a service for handling raw data provided by data providers. Further, it can also provide infrastructure (including GPU) as a service for training and testing of AI models.



# 03 DEPA's techno-legal approach to AI Governance

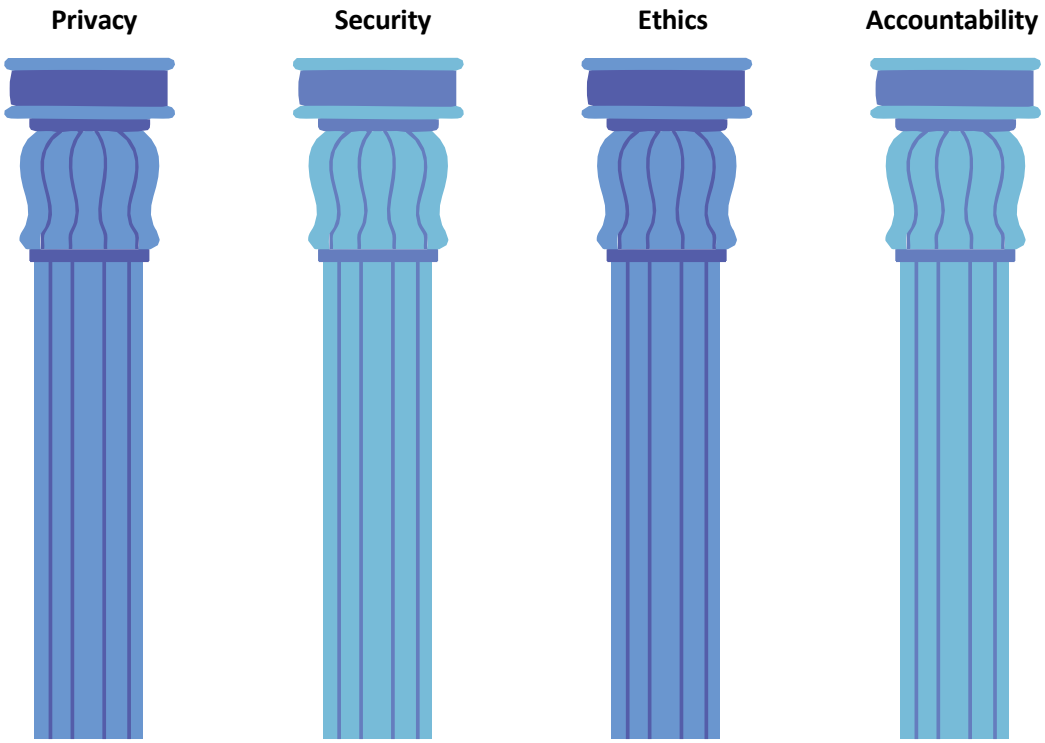
AI governance demands a fundamentally new approach—traditional static regulations simply cannot keep pace with this rapidly evolving, complex adaptive technology. India's solution leverages its proven expertise in building state-of-the-art Digital Public Infrastructures, adopting a techno-legal approach to governing data collaboration and AI innovation. Bringing technology artefacts of machine-readable electronic contracts, their cryptographic execution within secure data environment, and as per the allocated privacy budget, DEPA for Training aims to establish a governance framework that simultaneously accelerates in-novation and enforces regulation.

DEPA for AI Training proposes to unite all critical stakeholders — data providers, AI developers, government representatives, technology specialists, and domain experts — in creating a consortium of market players that can offer the necessary oversight for the functioning of a trustworthy and responsible innovation ecosystem.

DEPA for Training requires an oversight body to serve as the anchor for essential functions that must cover the entire lifecycle of AI development, including:

- Setting clear standards for privacy, ethics, and security,
- Registration of eligible participants,
- Overseeing techno-legal mechanisms for data sharing procedure and
- Facilitating dispute resolution.

By operating with sector-specific insights, ensuring alignment with foundational laws like the IT Act, 2000, DPDP Act 2023, and operationalizing AI governance through a techno-legal approach, DEPA for training offers a dynamic yet resilient structure necessary for responsible innovation.



# 04 Unleashing Sectoral Transformation

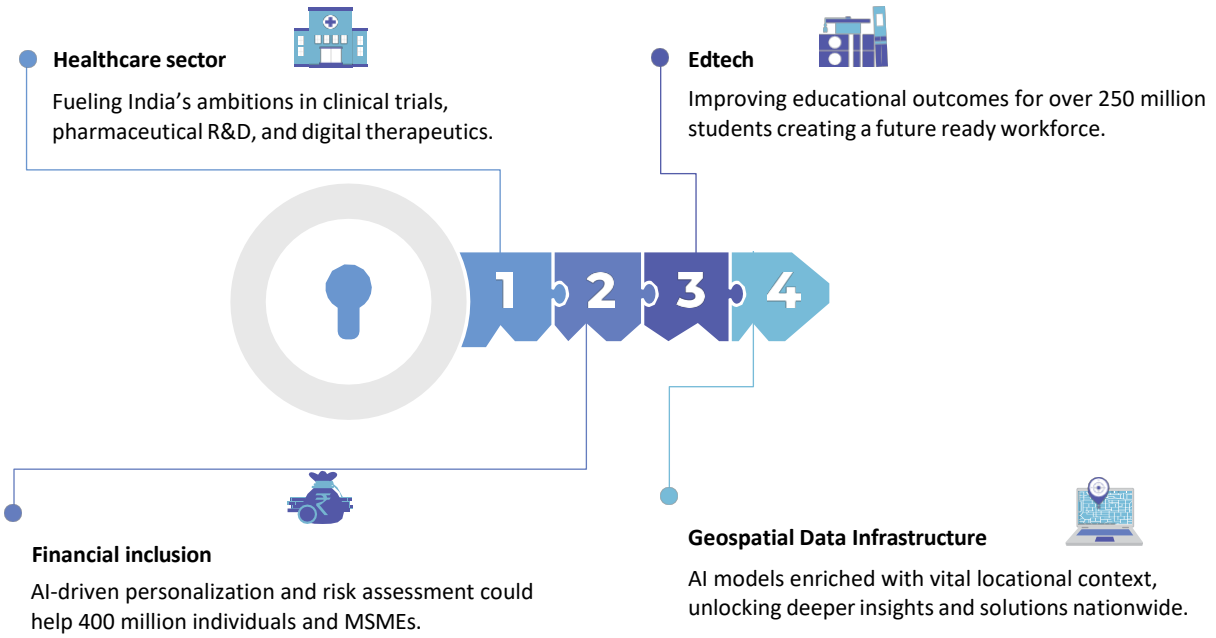
With DEPA Training architecture operationalized across sectors, India's DPI for AI stands poised to translate its unique digital advantages into leadership within the $15 trillion global AI economy. The framework provides the potential to unlock $500 billion in AI-driven value for India's GDP by 2030, moving beyond potential towards tangible economic and societal impact enabled by this new data-sharing paradigm.

The power of the DEPA framework is further amplified when combined with complementary national data exchanges such as the AI KOSH, API Setu, NIC API Exchange Platform (NAPIX), National Data and Analytics Platform (NDAP), Geospatial Data Infrastructure (GDI) and Open Government Data (OGD). By enabling secure access to the data exchange platforms (secure data environments), AI models across all sectors can be enriched with vital context, unlocking deeper insights and more effective solutions nationwide.

Implementation of DEPA promises transformative impact across a range of sectors. In Healthcare, it can unlock currently fragmented health data to power breakthroughs in AI-driven diagnostics, drug discovery, and public health management within a burgeoning $600 billion+ market. Within India's $550 billion Financial Services landscape, DEPA can enable secure sharing of diverse financial data, potentially extending formal credit to 400 million underserved individuals and MSMEs and fueling next-generation fintech innovation.

Similarly, in Education, DEPA can overcome data interoperability challenges to enable AI-powered personalized learning for millions and create portable skill credentials essential for a future-ready workforce of its 250 million students in a $29 billion EdTech sector. The successful adoption of DEPA Training across these domains is the essential catalyst for realizing India's vision of inclusive growth and responsible AI leadership on the global stage.

### India's DEPA for training is the key to unlocking massive potential across sectors

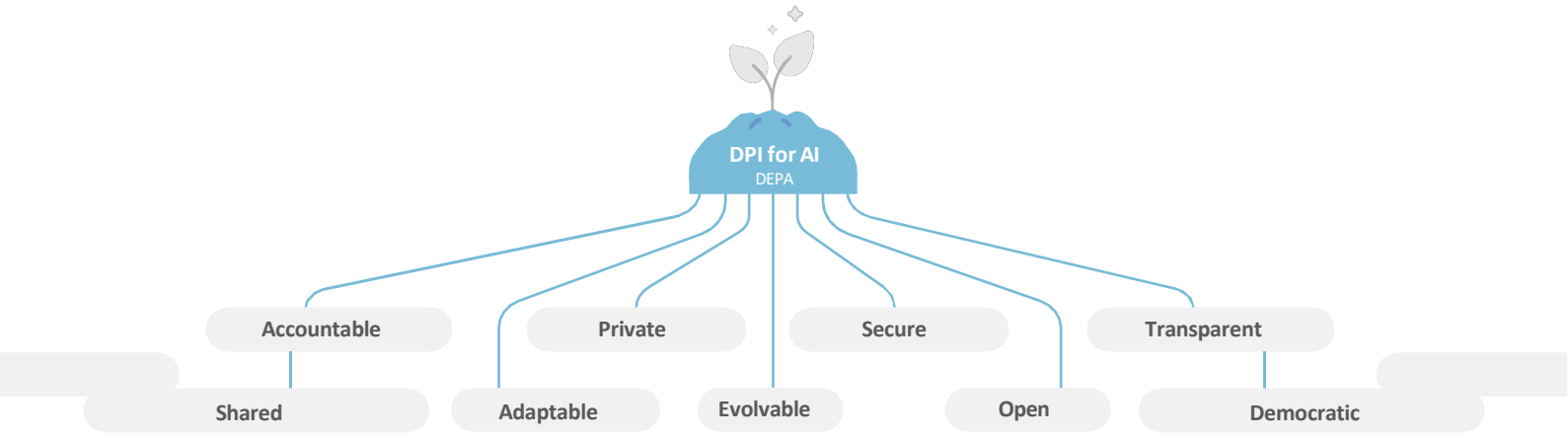


# 05 A Global Vision for Responsible innovation



Beyond national transformation, DEPA Training embodies India's commitment to responsible global leadership in the AI era. Its principles of ethical data sharing, privacy protection, and democratized access present a potential blueprint for global efforts in this direction, such as the UN AI Advisory Body's call for global training data framework (recommendation #6) to address training data challenges, fostering collaboration built on shared standards and mutual trust.

DEPA for Training is architected not merely as policy, but as a decentralized protocol – a digital public infrastructure enabling secure data collaboration by design. This protocol-based model holds the potential to empower nations worldwide, allowing them to train sovereign AI models on sensitive local datasets while also participating safely in vital cross-border research and development. By providing a common, trustworthy framework, DEPA can be the binding factor that allows a thousand AI models to bloom globally, addressing diverse challenges and advancing collective progress responsibly, actionizing the Indian philosophy of Vasudhaiva Kutumbakam – the world as one family.



# Table of Contents


01 The Data Deadlock ____ 17

02 DEPA for AI Training - Architecture Overview ____ 23

03 DEPA for AI Training - techno-legal approach to Governance ____ 30

04 The Market Opportunity of a Data-Fluent Economy ____ 35

# 1. The Data Deadlock

## 1.1 Understanding the Critical Role of Training Data

Artificial Intelligence (AI) is rapidly reshaping industries, economies, and societies worldwide. From powering recommendation engines and enabling medical diagnoses to optimizing supply chains and driving autonomous vehicles, AI's potential seems boundless. However, the engine driving this revolution, the very fuel that allows AI models to learn, adapt, and perform complex tasks, is **data**. Specifically, AI relies heavily on vast, diverse, and high-quality **training data.**

**What exactly is training data?**

It is the curated collection of information used to adapt or optimise an AI model. For an AI model, this data comprises numerous examples. In a supervised setting, each data point typically contains an input and the desired output or a label identifying the category of this data point. For instance, training an AI model to identify tumors in medical images requires showing it thousands of images, some labeled "tumor detected" and others "no tumor detected." Meanwhile, in an unsupervised setting, for example to build a financial fraud detection model, the AI model needs exposure to millions of transactions. Those that are outliers to the distribution may be classified by it as fraudulent or illegitimate.

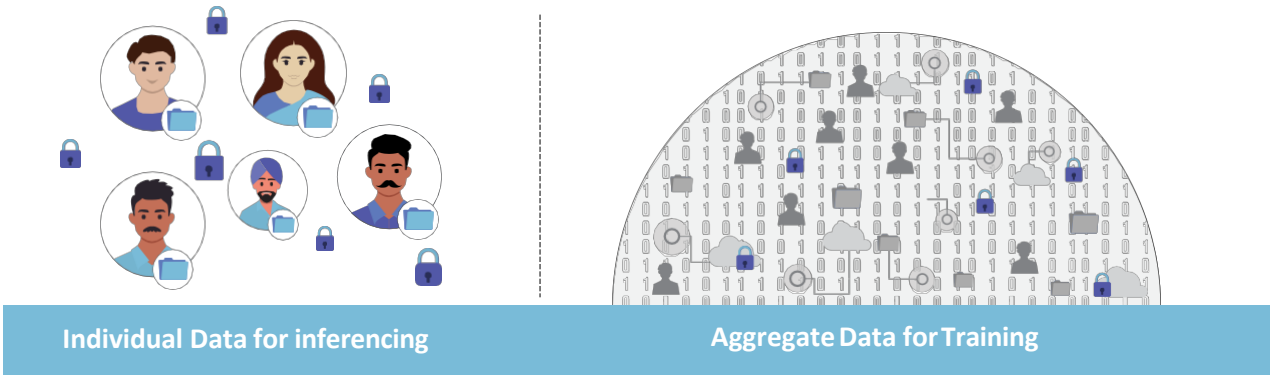
Importantly, a crucial characteristic of training data, especially for building sophisticated models like Large Language Models (LLMs) or complex predictive systems, is its **aggregate nature.** While it originates from individual data points (like a single transaction, a specific medical image, or a user's click), its power lies in the collective patterns, correlations, and statistical properties that emerge only when looking at the data en masse. An AI model learning about market trends doesn't focus on one person's purchase; it learns from the combined purchasing behaviour of millions. Similarly, a medical diagnostic AI model learns patterns from the aggregated health records of countless patients, not just one individual's history.

This is where the distinction between **training** and **inferencing** also becomes vital. **AI inference** is the process of applying pre-built intelligence i.e. an already trained model to make prediction or decision for an input data point that may be received individually. A clear example is the use of your personal banking data, accessed through a consent-based framework like DEPA's Account Aggregator, to get a pre-approved loan offer (or not) using an existing credit-scoring model. **AI training,** on the other hand, refers to the foundational process of building the AI

model's intelligence. This requires feeding the system vast amounts of data, allowing it to learn statistical patterns, correlations, and insights from collective information – the value lies in this broad statistical landscape painted by aggregate data. This is how granular details of one single individual when observed across a whole sub-population are rendered significant for the model.

In conclusion, training is about building intelligence using large-scale, aggregate data, while inferencing is about applying that intelligence to specific inputs or data points for decision making. Understanding this fundamental difference and complementarity – the need for broad, aggregate data for training versus individual data for inferencing – is the first step towards appreciating the unique challenges that DEPA Training addresses in making collective training data safely accessible for innovation.

**Training data is the curated collection of information used to adapt an AI model.**




## 1.2 Why is Valuable Training Data Inaccessible?

India stands at a unique juncture. With rapid digitization across sectors spurred by the digital ecosystems developed in-house, the country is generating data at an unprecedented scale. From healthcare records in hospitals and clinics, financial transactions flowing through banks and fintechs, educational data in schools and universities, agricultural data from farms to consumer behaviour patterns captured by e-commerce platforms – India is undeniably 'data-rich'.

If the data exists and holds such potential, why isn't data collaboration happening more readily for AI training? The barriers are significant and multifaceted, extending beyond simple technical hurdles.

**1.2.1** Firstly, **privacy risks** are a paramount concern. Even if data is intended for aggregate analysis, the raw data often contains Personally Identifiable Information (PII) or sensitive details. Using large datasets for training, even internally within an organization or with trusted partners,

increases the surface area for potential breaches or misuse. Accidental leaks or malicious attacks could expose highly sensitive information, leading to severe consequences for individuals and reputational damage for the data custodian. India's Digital Personal Data Protection Act (DPDP) 2023 rightly imposes strict obligations on handling personal data, making institutions cautious about any form of sharing that could risk non-compliance.

**1.2.2** Secondly, there are significant **institutional and business risks** beyond individual privacy. Businesses possess competitively sensitive information embedded within their datasets – customer lists, transaction patterns, operational metrics, proprietary algorithms. Using such training data, even if stripped of direct PII, could inadvertently reveal business secrets, intellectual property, or copyrighted material to competitors or other parties. The fear of losing a competitive advantage or enabling IP theft often outweighs any perceived benefit of data sharing for AI training.

**1.2.3** Thirdly, the lack **of trusted and accessible frameworks** exacerbates the problem. How can data collaboration happen securely? Who ensures it's used only for the agreed-upon purpose? How is compliance monitored and enforced? In the absence of standardized, trustworthy protocols, neutral intermediaries, and financially sustainable options, the perceived risks and complexity of sharing often seem insurmountable. The operational landscape of compliance for aggregate data specifically for AI training can be daunting, leaving institutions uncertain about permissible methods and potential liabilities.

**1.2.4** Finally, the **cost-benefit analysis** often tilts against data collaboration. Institutions bear the cost and risk of preparing, securing, and sharing data, while the benefits – often accruing to third parties (like AI model developers) – may not seem immediate or tangible enough to justify the effort and exposure. Without a clear return on investment or a way to mitigate risks effectively and make the process seamless and cost-efficient, the default position is often to keep data locked down. These combined factors create a formidable barrier, stifling the flow of data crucial for AI innovation.

## 1.3 A New Approach: Beyond Individual Consent

Recognizing the need for better data governance, India introduced the Data Empowerment and Protection Architecture (DEPA Inference), primarily operationalized through the Account Aggregator (AA) framework in the financial sector. DEPA Inference represents a significant step forward, empowering individuals with control over their personal data. Through Consent Managers (like AAs), individuals can grant explicit, granular, revocable consent for their data (e.g., bank statements, tax filings) to be shared securely from a Financial Information Provider

(FIP, like a bank) to a Financial Information User (FIU, like a lending institution) for a specific purpose, such as applying for a loan.

DEPA Inference excels at facilitating secure sharing of individual data, framing inference as a specific 1-on-1 transaction or service done in exchange for consent. It allows a small business to use its GST data to prove creditworthiness or an individual to share their mutual fund holdings with a new financial advisor. This consent-driven, individual-focused model is crucial for empowering users in their direct interactions with service providers.

However, the challenge of AI training data requires a different approach. While DEPA Inference manages the consented flow of individual data for inferencing (using an existing model), **it isn't designed to handle the privacy-preserving aggregation of bulk, de-identified data needed to build the AI models in the first place. The core issues are different.**

Consider a MarTech (Marketing Technology) use case. Under DEPA Inference, a user might consent to share their purchase history with Retailer A to Retailer B to receive a personalized discount coupon. This is individual data shared for a specific inferencing task (applying a discount rule). But what if Retailer A, B, and C want to collaborate to build a powerful AI model that predicts future purchasing trends across the entire market? They need to pool insights from millions of customer transactions without revealing individual identities, violating privacy, or leaking commercially sensitive sales figures to each other. Simple anonymization won't work due to re-identification risks and utility loss. In other words, individual consent via DEPA Inference is impractical at this scale and doesn't solve the institutional risk problem.

This is the gap: DEPA Inference addresses consented sharing for individual benefit in specific transactions; it doesn't provide the framework for secure, privacy-preserving pooling of aggregate data required for building powerful AI models. This necessitates **DEPA for AI Training** (and **Data Sharing for AI Training Framework),** a new layer designed specifically to unlock the value of aggregate data for AI training while upholding privacy and security through advanced technologies and governance mechanisms.

## 1.4 Risks of Inaction

In this age of AI innovation, it is imperative upon economies to address the challenges of accessing new and complementary pools of data for training AI models. Valuable data sets need to be made available for secure and private linking, enabling AI models to be trained upon connected, aggregate datasets for combinatorial problem solving. In the absence of such a framework, downsides are significant.

First, **innovation would be stifled.** Without access to diverse, large-scale datasets, Indian startups, researchers, and even established companies outside the dominant tech giants would struggle to develop cutting-edge AI solutions tailored to local needs.

Second, the **risk of privacy violation and data misuse would increase.** In the absence of a robust, secure, and trusted framework, pressure to access data may lead institutions to resort to ad-hoc methods. This could increase the likelihood of unauthorized use of personal information, and erosion of public trust.

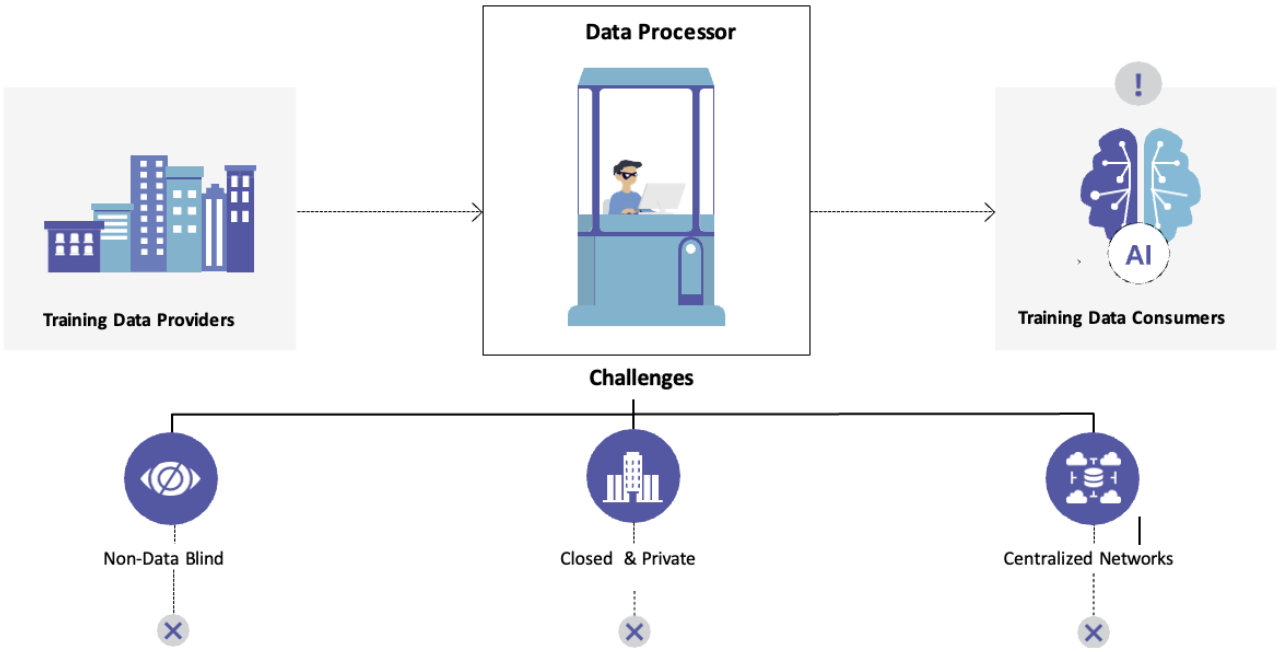
Third, the current situation could lead to **market concentration, and disadvantage local players.** Without a trusted and secure data collaboration framework there might be a create of an uneven playing field, where innovation becomes concentrated in the hands of a few. India possesses a key strategic advantage in its vast data generation; it is thus a national imperative to leverage it responsibly to gain new and impactful ground in the global AI race.

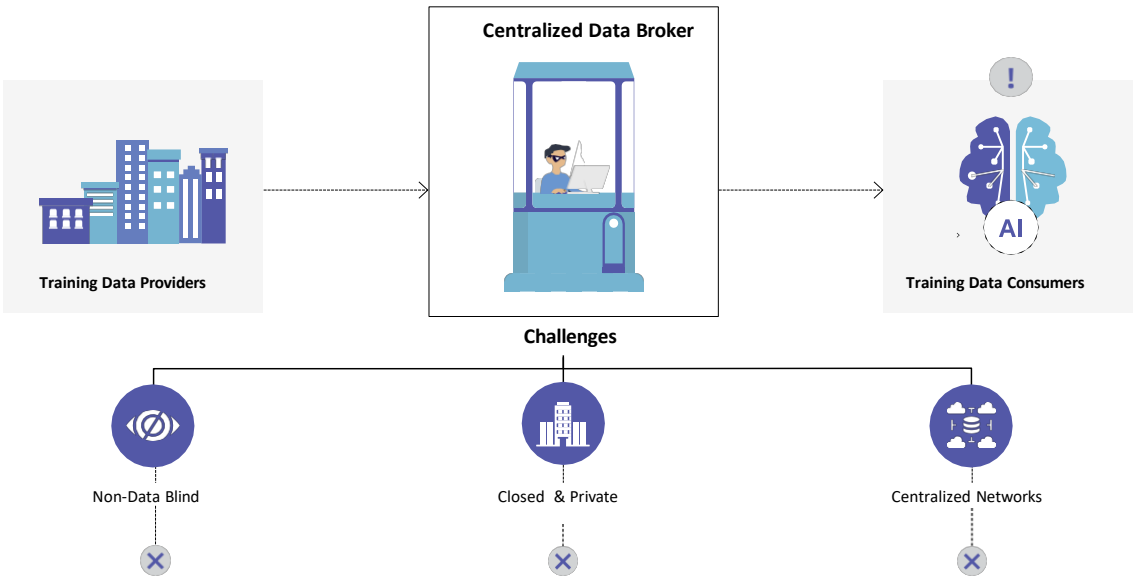
Compounding these structural issues are persistent **privacy vulnerabilities and unresolved intellectual property conflicts,** which collectively create a data deadlock, keeping valuable information locked in silos.

To conclude, inaction is not a neutral stance; it actively allows risks to grow and opportunities to wither. Establishing a framework like DEPA Training is not just about enabling AI; it's about mitigating these substantial risks, build a trustworthy and decentralized DPI to ensure that India can harness the power of data for inclusive innovation, economic competitiveness, and the benefit of its citizens, all while upholding stringent privacy standards.

## The current data architecture is broken

| Challenges with Data Providers | Challenges with Data Processors | Challenges with Data Consumers |
|---|---|---|
| • Missing data sharing procedure | • Lack of Sandbox | • No visibility of data usage |
| • Lack of MDDS | • Missing check for Anonymization | • Unseen IP / Patent generation |
| • No SLA | • API change without information | • No accountability to data principal or Data Providers |



# 2. DEPA For AI Training – Architecture Overview

## 2.1 A Framework for Trustworthy Collaboration

This chapter details the architecture of that solution: DEPA Training. DEPA Training is our framework Digital Public Infrastructure (DPI) for AI: a structured, privacy-centric ecosystem designed to facilitate secure, private, compliant, and ethical AI model training. It extends the philosophy of DEPA Inference, which empowered consent-based, individual-centric data sharing for inference, to the more complex challenge of enabling collective data use for AI model training. The architecture is an application of the foundational DEPA ADEPTS principles—Accountability, Democratization, Evolvability, Privacy, Transparency, and Security—to the unique context of ma chine learning with aggregate data.

At its core, frameworks like Data Sharing Procedure and DEPA for AI Training are built on a federated, decentralized architecture. It moves away from opaque bilateral agreements and the risks of centralized data brokers, instead establishing a governed space for collaboration between data providers and AI model developers animated by a range of enabling participants (cloud service providers, model providers, certifiers, etc). This collaboration is mediated through sector-specific governance and privacy-preserving technologies implemented through Data Collaboratives etc., ensuring that confidentiality and compliance are maintained throughout the AI lifecycle. By integrating Secure Data Environments and Differential Privacy, the framework allows multiple actors to contribute and utilize deidentified aggregate data responsibly, without direct exposure or transfer of raw information.

What follows is a detailed outline of the DEPA training's structure and workflow. We will describe the roles and responsibilities of its key participants, from the institutions that provide data to the innovators who consume it, to new and enabling technology service providers and the governance bodies that oversee their interactions. We will then elaborate on the underlying technical components that enforce policy, ensure traceability, and deliver end-to-end data protection. This techno-legal framework establishes a model for AI development that upholds trust by law, trust by design, and trust by technology.

## 2.2 Key Players of the DEPA Ecosystem

DEPA for Training is composed of distinct but interdependent participants, each with a defined role. This structured separation of duties brings clarity, accountability, and auditability to a process that has historically been opaque.

**Data Principals**

At the very foundation of the ecosystem are the Data Principals, the individuals or entities to whom the data originally belongs. While they may not interact directly with the training network, their rights are the bedrock upon which the entire architecture is built. Their data is managed under legal regulations that mandate purpose limitation and privacy safeguards, and their interests are protected through the institutional governance structure that operationalises cutting-edge technology – electronic contracts, confidential computing, and differential privacy, to this end.

**Data Collabs**

Bridging the gap between data principals and the AI innovation network of modelers is a new class of actors: Data Collabs or collectives. These intermediary institutions, which can be cooperatives, industry associations, or research consortia, work with communities to aggregate, standardize, and organize data from multiple sources. They play a vital role in synchronising fragmented data streams for further processing and may be responsible for obtaining primary consent where applicable. This layered approach allows the network to scale and incorporate data from sources often underrepresented and complementary in mainstream Al development.

**Training Data Providers (TDPs)**

Training Data Providers are the formal gatekeepers of data within the DEPA ecosystem. They are institutions responsible for sourcing, curating, and preparing datasets for AI training. A TDP's role extends beyond simple custody; they must apply privacy-preserving transformations and ensure datasets meet stringent technical and regulatory requirements for deidentification. TDPs can come from any sector and may include, but not be limited to:

- Hospitals and healthcare providers contributing anonymized patient records for predictive health modeling.
- Financial institutions sharing deidentified transaction data to improve fraud detection or credit risk models.
- Public agencies and smart city operators offering data streams for urban planning and traffic optimization.
- Pharmaceutical companies sharing data with smaller firms for drug discovery and repurposing initiatives.

TDPs are registered with the central registry and are bound by its governance standards, which are enforced through electronic contracts.

**Training Data Consumers (TDCs)**

Training Data Consumers are the innovators who access these datasets to develop, refine, or evaluate Al models. TDCs must operate within strict boundaries of purpose limitation and ethical conduct defined in their data-sharing agreements with Data Collectives federating data sets for a given sector from multiple TDPs or with individual TDPs directly. The network is designed to fuel innovation across a spectrum of TDCs, including:

- Academic and non-profit research institutions developing open-source models for the public good.
- Start-ups and enterprises, who are often data-hungry but low on resources, building AI-driven commercial solutions.
- Regulatory bodies seeking to test the fairness, safety, and accuracy of third-party Al systems in sensitive domains.

All model training by TDCs must occur within approved secure data environments, and the resulting models are subject to validation.

## 2.3 Privacy and Security by Design

The integrity of the DEPA Training Framework rests on a technological foundation where security and privacy are embedded into the architecture itself. This “secure by design” approach is essential for building the institutional and public trust required for large-scale data collaboration. Two pillars form the backbone of DEPA’s technical guarantees: the rigorous application of Differential Privacy and the comprehensive enforcement of a Zero-Trust Security Framework.

### 2.3.1 Why Traditional “Anonymization” Isn’t Sufficient for training AI models

A common suggestion for overcoming data collaboration barriers is “anonymization” – simply removing direct identifiers like names, address- es, or unique ID numbers. The idea is that once PII is stripped, the data becomes non-personal and can be used for training freely without privacy risks. Unfortunately, for the purposes of generating robust AI training data, this approach is often a dangerous oversimplification – a myth that fails to grasp the complexities of modern data.

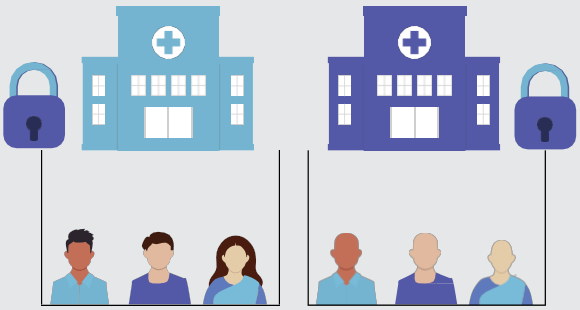
The core problem lies in **reidentification risk.** Decades of research have shown that even data stripped of direct identifiers can often be re-identified by linking it with other available datasets. Seemingly innocuous pieces of information (like date of birth, zip code, and gender; or a combination of transaction timestamps and locations) can, when combined, act as a unique "fingerprint" identifying an individual. As datasets grow larger and more diverse, and as auxiliary datasets (from public records, social media, data brokers) become more readily available, the possibility of re-identifying individuals within supposedly "anonymized" datasets increases dramatically. Combining just two or three moderately anonymized datasets can exponentially increase the risk, potentially revealing sensitive information about individuals thought to be protected.

Furthermore, the very process of rigorous anonymization required to significantly reduce reidentification risk often compromises the **utility** of the data for AI training. Techniques like k-anonymity, l-diversity, or adding significant statistical "noise" involve removing or generalizing data points. While this enhances privacy, it can distort the underlying patterns and correlations that AI models need to learn effectively. Overly anonymized data might prevent reidentification but result in poorly performing, inaccurate, or biased AI models. There is a fundamental trade-off: aggressively anonymise and lose data utility; preserve utility, and risk reidentification.

**Why India's Vital Health Training Data is Siloed**



**Vast Health Datasets Exist Across India, But Remain Locked in Institutional Silos, Halting AI Progress**



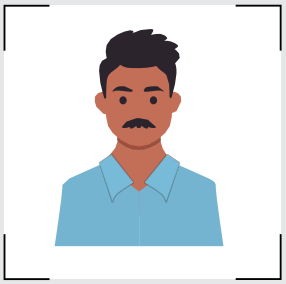
Valuable Health Data

**Why Anonymisation Fails**



Record 1 (Seemingly Anonymous) + Anonymized Dataset A (e.g., Age + Diagnosis) + Anonymized Dataset B (e.g., Pincode + Hospital Visit)

Combining Multiple 'Anonymous' Data Sources



Can Re-Identify Individuals, Exposing Sensitive Profiles!



Privacy Violation | Risk of Harm | Loss of Business Confidentiality | Security & IP RiskSecurity

Traditional anonymization techniques were often developed when datasets were smaller and computational power was limited. They are frequently inadequate for the massive, high-dimensional datasets required by modern AI and the sophisticated techniques used to analyze them. Relying solely on stripping direct identifiers provides a false sense of security and fails to address the nuanced privacy challenges inherent in sharing data for complex AI model training. A more sophisticated approach is clearly needed.

### 2.3.2 Differential Privacy Mechanisms

DEPA confronts this challenge with Differential Privacy, a mathematically robust technique that allows for learning from aggregate trends while making it statistically improbable to infer information about any single individual.

Instead of simply stripping identifiers, this method introduces carefully calibrated statistical noise into query responses or model outputs. This ensures that the inclusion or exclusion of any single person's data has a negligible effect on the result, thus severing the link between the aggregate insight and the individual contribution. This is crucial in high-stakes domains like healthcare, where de-identified patient records can be used to train diagnostic models without compromising patient confidentiality.

Within DEPA, the implementation of differential privacy is governed and audited. Privacy budgets—which limit the cumulative privacy loss from repeated queries are transparently defined and monitored, allowing for a managed trade-off between privacy and analytical accuracy tailored to the risk profile of each sector.

### 2.3.3. Zero-Trust Security Framework

Complementing data-level privacy is a network-level security posture governed by a Zero-Trust Security Framework. This model discards the outdated notion of a trusted internal network, operating instead on the principle of "never trust, always verify". Every interaction within the DEPA ecosystem is treated as potentially hostile and must be explicitly authenticated and authorized. In practice, this means:

- **Strict Verification:** Every data access request, regardless of its origin, is subjected to rigorous verification of identity, role, and policy compliance before access is granted.

- **Pervasive Encryption:** Data is protected at all stages—at rest, in transit, and during processing—using mandatory, industry-standard encryption protocols.

- **Continuous Auditing:** AI models and technology environments like secure data environment etc undergo regular security audits to detect vulnerabilities and enforce access policies.

- **Real-Time Monitoring:** Access logs are maintained in real-time, and any anomalous activity, such as an unauthorized access attempt, automatically triggers an alert and investigation.

This continuous enforcement, combined with cryptographically signed contracts and datasets, creates a closed-loop system of accountability. Together, differential privacy and zero-trust architecture ensure that AI training within DEPA is accelerated whilst also proceeding within legal and ethical bounds, and in a manner that is fundamentally secure by design.

### 2.3.4 Data Exchange Platforms and Secure Data Environments

Data exchange platforms and secure data environments are the technological cornerstone of DEPA's privacy infrastructure. They are specialized, high-assurance computational environments where AI models can be trained on sensitive, anonymized data without ever exposing the raw data to TDCs or other external systems. This eliminates the need to transfer or expose raw data, minimizing the risk of breaches or unauthorized use. Key safeguards embedded within secure data environment include:

- Execution of all computations within secure data environments, which are secure hardware enclaves that prevent data leakage or tampering during processing.
- The application of privacy algorithms like Differential Privacy to ensure individual patterns cannot be inferred from model outputs.
- Support for advanced techniques like Federated Learning and Secure Multi-Party Computation (MPC), which allow for collaborative training across different data silos without pooling the data itself.

Providers of secure data environments etc must be registered, ensuring they meet rigorous security and privacy standards.

## 2.4 Operationalising DEPA Training

### 2.4.1 Workflow – Data Sharing Procedure

Ministries, Departments, and Organisations (MDOs) should aim to publish a Data Sharing Procedure specifying dataset classification (open, registered, restricted), consent mechanisms, de-identification standards, API access processes, and penalties for misuse. Personal and sensitive data shall not be shared in raw form; de-identification with pseudonymized keys held by the originating agency is mandatory before sharing. All shareable documents viz.

anonymized personal, business, legislative, or administrative data shall be available in machine-readable format through APIs.

The data sharing procedure for AI training would additionally require support such as:- DEPA Training functions through a meticulously orchestrated workflow that ensures security, compliance, and accountability at every step. This sequence transforms raw, distributed data into ethical and certified Al models, governed by clear rules and technical safeguards.

**Step 1: Data Registration and Encryption**

The process begins when a Training Data Provider (TDP) formally registers a dataset with the DEPA network which may be hosted on Data Exchange Platforms like AI Kosh directly through self-hosted catalog services or through a Data Discovery Agent or a Data Collective – an action overseen by a governance body. Before any data can be used, it must be prepared and catalogued. Prior to this stage, datasets are subjected to anonymization and encryption to remove personally identifiable information by data providers. Each dataset is assigned structured meta data describing its origin, sensitivity, and access restrictions, and is cryptographically tagged or watermarked to ensure end-to-end traceability. This registered and protected dataset is now discoverable by authorized data consumers at the Data Exchange Platform through the DEPA framework.

**Step 2: Data Request and Contract Validation**

Next, a Training Data Consumer (TDC) initiates a request to access one or more datasets for a specific, use case -. This request is not a simple query; it is a formal proposal that is validated by the DEPA Electronic Contract Management System. The system automatically checks the TDC's request against the data-sharing policies defined by the Data Collective or the TDP and the governance rules. This automated enforcement of contractual terms ensures that data is only accessed by authorized parties for legitimate and registered purposes.

**Step 3: Secure Model Training in Data Exchange Platform and Secure Data Environments**

Upon successful validation, data aggregation and analysis is provisioned exclusively within secure data environment. The platform design enforces a strict no-exposure policy for raw data. Using hardware-based Secure Data Environments, the data remains encrypted and isolated even during processing. Within this secure environment, Differential Privacy techniques are systematically applied to prevent the model from learning or revealing individual specific information. All activities inside the secure data environments are immutably logged for audit purposes, ensuring that the process is both secure and transparent.

**Step 4: Model Validation and Bias Detection**

Once a model has been trained, it must undergo a rigorous, independent validation process. This assessment is multifaceted, covering:

- **Bias and Fairness Audits:** The model is tested to ensure it does not produce discriminatory outcomes or amplify historical inequities.

- **Security and Robustness Testing:** The model is evaluated for vulnerabilities to adversarial attacks or potential data leakage.
- **Large-Scale Safety Testing:** For high-risk applications, models undergo additional testing, including complex adaptive systems modeling and adversarial black-box testing.
- **Algorithmic Explainability:** The model's logic is assessed to ensure its decisions can be interpreted and explained in legal and practical contexts.

Only models that pass this comprehensive check receive go ahead for pilot deployment and scaling.
This check will be as per the guidelines issued by the competent government authority.

**Step 5: Deployment and Continuous Monitoring**

The TDC can deploy the AI model for public or commercial use. However, the oversight does not end at deployment. The DEPA framework requires continuous monitoring to ensure sustained compliance and safety. This includes maintaining fairness and security logs for periodic review and ensuring full traceability from a model's output back to the data and contracts involved in its training. This final step ensures long-term trust and accountability for all AI systems developed within the DEPA ecosystem.

### 2.4.2 Economic & Business Model

DEPA Data Collaboratives (Collabs) serve as nodal entities bringing together various market stakeholders onto one common platform. In that, Data Collabs are business centric, sector specific and conceptualized to be commercial players. They exist to break silos and bring consensus, efficiencies and measurable impact to the task of data sharing and subsequently, building of AI models that are safe and responsible.

Data Collabs' business model involves providing critical ecosystem services and tools including,

- General tools and services (registration, authentication, curation for discoverability)
- Data-related tools and services (standard schemas, encryption-decryption protocol, exchange protocol)
- Domain related tools and services (e-contract)

There are built in incentives for all the other ecosystem actors and enablers. Through the platform, data providers (TDPs) can capitalize on their data repositories, entering a benefit sharing model whereby for example hospitals that share data are able to get priority use of models that emerge from the data to improve clinical and patient care outcomes.

For data modelers, their innovation advances stem directly from the access to bigger and better-quality data. Other players such as Technology service providers (Compute providers, Model providers) can align, refine and position their product and services to the emerging needs of the Data and AI innovation sector.

# 3. DEPA techno-legal approach to Governance

## 3.1 A New Model for AI Governance

AI is a complex, and rapidly evolving technology. Static, top-down regulations, enforced through manual screening or audits are ill-suited to the pace of this digital technology. On one hand, they risk becoming a bottleneck to innovation, and on the other becoming ineffective safeguards due to enforcement challenges. Where DEPA for inference successfully leveraged a centralized regulatory body like Sahamati to orchestrate the Account Aggregator ecosystem, DEPA for Training offers greater challenges. It must enable multiparty data sharing whilst guaranteeing data use within prespecified security and privacy provisions. Fair and transparent use of data are equally important guarantees for responsible AI innovation. DEPA for training is thus conceived upon the ADEPTS principles, with the core ideals of Accountability, Democracy, Evolvability, Privacy, Transparency, and Security built into the framework by design, and operationalized through the novel approach of Techno-legal governance.

## 3.2 Embedding Compliance by Techno-legal framework

The DEPA for Training framework's strength lies in its fusion of legal principles with technological enforcement. Compliance is not left to the goodwill of the ecosystem; it is woven into the very architecture of the system. It moves beyond policy to operationalize trust through a cohesive techno-legal framework that provides both robust process guarantees and verifiable computational guarantees.

This "compliance by design" approach offers a direct and concrete set of solutions to the enforcement deficit, ensuring that privacy and ethical considerations are protected not just by policy, but by design and by technology.

DEPA's computational guarantees provide the hard, technical enforcement of its core principles. The primary guarantee is the use of secure data environment for AI model training. By leveraging hardware-based secure data environment, Data Exchange Platform (like AI Kosh) create a secure enclave, wherein sensitive, anonymized data can be processed without ever being exposed in its raw form.

This is a direct, architectural solution to the risks of data breaches and unauthorized use, enforcing the principles of "privacy by design" and "data minimization" at the hardware level.

Complementing the security of Secure Data Environments is the mandated use of Differential Privacy. This mathematically robust technique provides a provable guarantee of anonymity, addressing the critical failure of traditional anonymization methods. By introducing precisely calibrated statistical noise, Differential Privacy makes it statistically improbable to infer information about any single individual from a model's output. Together, these process and computational guarantees create a closed loop system of accountability. In this model, compliance is not a barrier to innovation; it is the feature that unlocks it.

In summary, governance is enforced through a suite of technical mechanisms:

- **Electronic Contracts:** These machine-readable agreements automate the enforcement of data-sharing conditions, purpose limitations, and usage rights, creating a transparent and auditable chain of responsibility.

- **Secure Data Environment:** As the mandatory environments for all model training, the secure data environments where needed technologically prevent raw data exposure, enforcing data minimization by default.

- **Differential Privacy:** This technique provides a mathematical guarantee of anonymity, ensuring that model outputs cannot be used to re-identify individuals, thereby aligning with the core tenets of data protection law.

- **Zero-Trust Security:** The architecture operates on a "never trust, always verify" basis, with every interaction requiring cryptographic authentication and authorization, all of which is recorded in immutable audit trails.

This framework is built to align with key domestic and international regulations, including India's Digital Personal Data Protection Act (DPDP), 2023, the EU's General Data Protection Regulation (GDPR) and AI Act, and other global risk management frameworks. It also accommodates sector-specific laws like HIPAA in healthcare or Basel III in finance.

## 3.3 An Architecture of Collaborative Oversight

DEPA for AI Training requires an agile governing body responsible for managing the integrity and accountability of the ecosystem. It must function both as a policy steward and a technical administrator, ensuring that every participant operates within a harmonized legal, ethical, and procedural framework. Its structure needs to be a deliberate departure from centralized control, designed instead as a consortium constituted by the ecosystem participants themselves, including different critical voices such as:

- Training Data Providers (TDPs)
- Training Data Consumers (TDCs)
- Data Collectives (Data Collabs)
- Technology Service Providers (TSPs)
- Sectoral domain experts
- Civil society representatives

DEPA for Training has conceived of this multi-stakeholder composition of oversight to ensure that the governing policies of any DEPA framework remain grounded in practical reality and reflect a balance of interests. It is aimed at fostering co-ownership and a shared sense of responsibility, where the rules of the road are set by those who travel on them.

Any governance framework that is developed should be the anchor point for a comprehensive set of responsibilities spanning the entire AI training lifecycle.

- **Participant Registration and Identity Management:** maintaining a central registry of all participants in the DEPA ecosystem, including formally registering and verifying the identity of TDPs, TDCs, and Secure Data Environment, and issuing cryptographic credentials necessary for them to interact securely within the network. The latest anonymisation technique and quantum safe cryptography should be used, and the regular update be factored in.

- **Mandatory check and having authority in the loop:** registering key infrastructure providers and certifying AI models. Secure Data Environment should meet prescribed security and privacy standards before they can operate within the DEPA network. Additionally, trained AI models must be verified as “fit for use,” verifying too their compliance with standards for fairness, bias mitigation, and data security before they can be deployed. This due diligence should be taken care by the competent government authority.

- **Managing Electronic Contract Infrastructure:** electronic contract system constitutes the operational backbone of the DEPA network. The contract can be managed and maintained electronically by the Institutional Governance Structure or Data Exchange Platform itself and/or the contract chain can be linked to "AI provenance Chain", a distributed ledger that provides an immutable record of the data processing operations that have proceeded on the data undersigned contractual agreements. This ensures full traceability, verifiability, and accountability of contractual terms.

- **Auditing and Compliance Monitoring:** Regular check of AI models and training workflows is essential to ensure adherence to DEPA's principles. This includes the power vested with the competent government authority to suspend or revoke non-compliant activities, providing a critical enforcement mechanism to maintain the integrity of the ecosystem.

- **Conflict Resolution:** In the event of disputes between participants such as disagreements over data usage, contractual breaches, or ethical concerns, formal mechanisms for mediation and arbitration would need to be defined, ensuring fair and efficient resolution.

By embedding these administrative and regulatory functions into the technical substrate of the DEPA system, DEPA for AI Training remains lightweight, decentralized, and legally resilient.

## 3.4 A Roadmap for Implementation

The successful adoption of this ambitious governance framework requires a structured, phased approach. To move from blueprint to a functioning ecosystem, the following roadmap is proposed:

1. **Establish Pilots:** The initial phase will focus on establishing pilot Data Exchange Platforms and Data Collabs, institutional governance structure, etc in a few high-impact sectors like finance and healthcare. This will allow for the refinement of governance protocols in real-world conditions.

2. **Implement Regulatory Sandboxes:** In parallel, regulatory sandboxes will be created to allow developers and institutions to train, test and validate AI models on DEPA datasets in a controlled compliance environment.

3. **Promote Cross-Border Governance Frameworks:** The Data Exchange Platform providers and other organizations and institutions will actively work to establish bilateral and multilateral agreements to facilitate secure international data collaboration and research, ensuring interoperability with global standards.

4. **Launch Awareness and Capacity-Building Programs:** A concerted effort will be made to educate developers, institutions, and policymakers on the principles and practicalities of the DEPA framework through targeted campaigns and workshops.

5. **Develop Sustainable Incentive Models:** To encourage long-term participation from data providers, the framework will explore viable incentive structures, such as royalty or access-fee models, that reward contributions to the data ecosystem.

Through this deliberate and collaborative rollout, the regulatory framework for DEPA training for Training can be established as a robust, evolvable, and trusted foundation for India's AI future, one that democratizes access to data while championing ethical and responsible innovation on a global stage.

# 4. The Market Opportunity of a Data-Fluent Economy

## 4.1 Quantifying the Cost of the Training Data Deadlock

Data silos—repositories of information isolated from an economy's broader architecture impose staggering costs. Globally, this fragmentation translates into trillions of dollars in unrealized revenue and productivity. This issue is particularly acute in India, where the vast digital footprint of its population remains largely inaccessible for domestic innovation, creating a critical data gap for startups, researchers, and policymakers.

The technical problem of data silos directly creates the economic problem of information asymmetry. This condition, where one party in a transaction possesses more or better information than the other, is a primary driver of market failure in developing economies. In India's credit market, this asymmetry manifests as a persistent and massive credit gap for Micro, Small, and Medium Enterprises (MSMEs). Lacking reliable data on new or small-scale borrowers, lenders rationally favor established entities with long credit histories. An expert committee constituted by the Reserve Bank of India (RBI) estimated this MSME credit gap to be in the range of ₹20 to ₹25 trillion.

India's existing Digital Public Infrastructure (DPI) has already proven its capacity to correct such market failures. The Unified Payments Interface (UPI) and Aadhaar dramatically reduced transaction costs and information asymmetry, enabling a level of financial inclusion in just a few years that would have otherwise taken nearly five decades to achieve. This provides a powerful precedent. By establishing secure, interoperable pathways for sharing aggregate, deidentified data, the DEPA for Training framework acts as a targeted DPI to resolve the information asymmetry at the heart of the training data deadlock, unlocking economic potential across the entire economy.

## 4.2 India's Strategic Position in the Global AI Economy

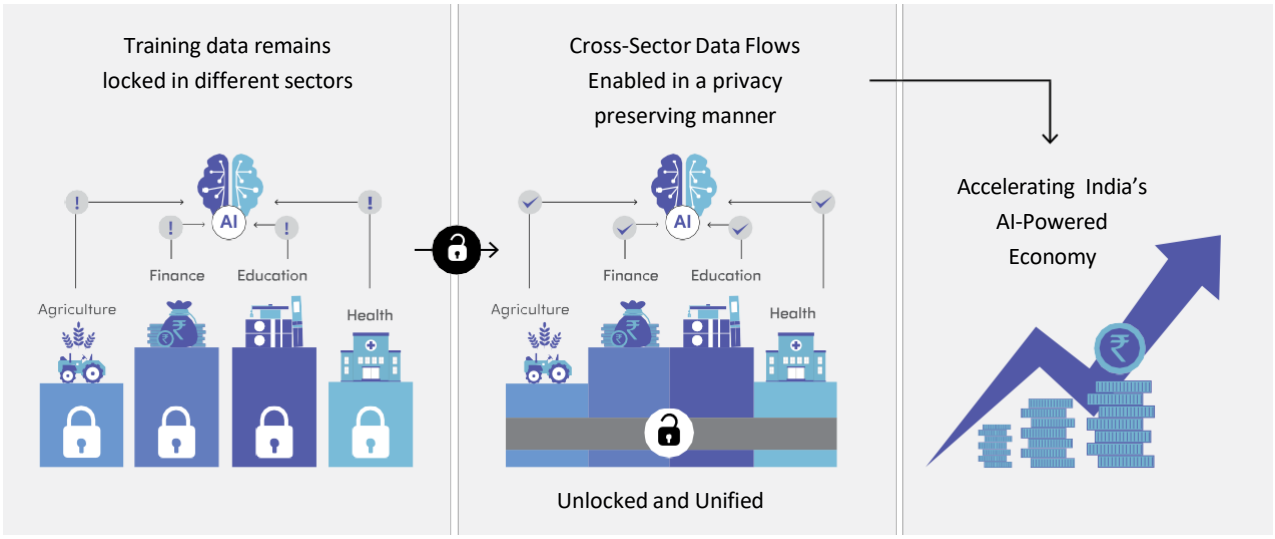
The global AI economy is expected to add over $15 trillion to global GDP by 2030, a transformation powered by data liquidity. India, with its billion-plus people online and an advanced stack of DPI is uniquely positioned to capture a significant share of this value. Projections indicate that India's digital economy will reach $1 trillion by 2030, with AI alone poised to contribute $500 billion to the nation's GDP.

However, realizing this potential requires moving beyond digitization alone. The next wave of growth will be driven by the ability to act on data in a trusted, secure, and value-generating way. The true power of data is realized not in isolation, but in its integration. This is where the synergy between DEPA for Training and complementary national infrastructure like the Geospatial Data Infrastructure (GDI) becomes strategically vital.

The GDI provides a unified catalogue of high-quality spatial datasets—from land-use layers to transport networks and satellite imagery—via open APIs. By enabling AI models to be enriched with this precise locational context in addition to sectoral datasets through the secure DEPA framework, a multiplier effect is created. Integrating sectoral data with a geospatial layer yields significant efficiency gains and enables entirely new forms of value creation. An AI model trained on health data can identify disease patterns; an AI model trained on health data combined with GDI data can predict a disease outbreak's geographic spread and guide resource allocation with pinpoint accuracy. This ability to create sophisticated, context-aware models is central to India's strategic opportunity.

**DEPA for AI Training**
**Data Sharing via Data Empowerment and Protection Architecture allows for data collaboration**

Empowering AI with Richer, Combined Insights



## 4.3 Sectoral Deep Dive

DEPA for AI Training is the key to unlocking tangible value in India's most critical sectors. The following analysis explores the specific market opportunities and the systemic barriers that the framework is designed to overcome.

**Healthcare and Life Sciences: From Fragmented Data to Precision Public Health**

India's healthcare sector, projected to reach $638 billion by 2025, stands at a critical juncture. It is characterized by pockets of world-class excellence existing alongside systemic fragmentation. This fragmentation imposes direct and quantifiable costs. While a comprehensive India-specific study on this is pending, international evidence suggests that fragmented care can nearly double healthcare costs. In the Indian context, a report by NITI Aayog estimates that reducing health system fragmentation could prevent 1.5 million households from falling into poverty each year due to catastrophic health expenditures.

The DEPA for Training framework offers a pathway to move beyond improving individual diagnostics toward creating population-level predictive intelligence. By enabling secure access to large-scale, aggregated health data, it opens new frontiers in:

- **Precision Medicine and Genomics:** The genetic makeup of the Indian population is unique. Training AI models on aggregated, pan-India genomic data can lead to the development of Polygenic Risk Score (PRS) models tailored to the Indian populace, enabling personalized disease prevention strategies at a massive scale.

- **Predictive Epidemiology:** AI excels at identifying patterns in complex data, making it a powerful tool for building early warning systems for disease outbreaks. DEPA for Training provides the mechanism to connect disparate data streams from platforms like the Nikshay portal for Tuberculosis management and the e-Sanjeevani telemedicine platform, creating a truly national predictive health intelligence system.

- **Drug Discovery and Development:** The framework can accelerate the development of novel drugs and the repurposing of existing ones for diseases prevalent in India by allowing researchers to train AI models on a rich combination of clinical trial data, real-world patient data, and genomic information.

The technical viability of this vision has been proven. The landmark Penn Medicine/Intel project, a federation of 71 international institutions, used Federated Learning within a Confidential Clean Room to train a superior AI model for brain tumor detection without sharing raw patient data.

The collaborative model achieved over 99% of the accuracy of a hypothetical model trained on centralized data. This case proves that privacy can be preserved with negligible loss of performance, but it also highlights that the greatest barriers are non-technical. The success of DEPA's Institutional Governance Structure for healthcare will hinge on its ability to provide the legal and governance clarity necessary to foster trust and participation.

**Financial Services: Unlocking Credit and Mitigating Risk**

India's financial services sector has made enormous strides in digital inclusion, yet an estimated 400 million individuals and a vast number of MSMEs remain outside the formal credit ecosystem. The "Training Data Deadlock" is a core reason for this impasse. The DEPA for Training framework is positioned to break this deadlock by enabling the secure, aggregated use of "alternate data," thereby unlocking credit for the underserved.

Traditional credit scoring models inherently exclude those without a formal loan history. Alternate data—such as transactional data from GST filings and e-commerce platforms, utility bill payment histories, and supply chain data—provides powerful indicators of financial health and repayment capacity. By enabling lenders and fintech innovators to train AI models on aggregated, deidentified versions of these datasets, DEPA can help create credit scores for millions of "credit invisible" Indians.

This directly addresses the root cause of the ₹20-25 trillion MSME credit gap: information asymmetry. By creating a trusted mechanism to pool and analyze aggregate transactional data, the framework reduces the perceived risk associated with new and small borrowers, allowing lenders to expand the credit market. This would not only fuel the growth of the MSME sector, which contributes nearly a third of India's GDP, but also enhance the stability of the lending institutions themselves by opening a vast new customer base.

The challenge is not merely a lack of will, but systemic friction. The RBI's recent regulatory action against major banks for deficiencies in their IT infrastructure highlights a core problem: many incumbent institutions are burdened with outdated legacy systems that are technically incapable of supporting modern, secure data environment. DEPA for Training provides a crucial "bypass" to this issue, creating a set of standardized, secure pipes through which aggregated data can flow, allowing innovation to proceed without being held hostage by the technological debt of individual institutions.

**Education and Skilling: Building a Future-Ready Workforce**

As India navigates the Fourth Industrial Revolution, the quality of its human capital will be its most critical asset. The education and skilling sector, projected to exceed $29 billion by 2030, faces the

task of preparing 250 million students for a future where skills, not just degrees, are the currency of employment. The DEPA for Training framework can serve as the foundational data infrastructure to enable this transition.

- **Adaptive Learning at Scale:** AI-driven adaptive learning systems can personalize education to individual student needs, a method shown in studies to improve test scores by an average of 30%. However, training these systems requires large, diverse datasets of learner performance, which are currently locked within institutional silos. DEPA can unlock this data, enabling the development of effective personalized learning platforms at a national scale.

- **The "Skills Passport" Dividend:** A critical failure of the current ecosystem is the inefficient signaling of skills between education and employment. Portable, verifiable "skill passports," or Learning and Employment Records (LERs), are a data-driven solution. By creating a trusted, interoperable system for recording and sharing verified credentials, DEPA can enhance labor market efficiency. Critically, DEPA can synergise with latest national digital registries such as Academic Bank of Credits (ABC) rolled out under the National Education Policy, 2020 to serve as a living bank account for academic currencies. Solutions can be tailored to help employers make more precise, data-driven hiring decisions, while workers are empowered with a comprehensive record of their lifelong learning, improving their mobility. This is particularly crucial for promoting equity, as it allows individuals to receive formal recognition for skills acquired through non-traditional pathways.

The experiences of other large, federated countries show that building such a national data framework is as much a political and organizational challenge as it is a technical one. The success of DEPA's Institutional Governance Structure for education will depend on its ability to build consensus on common data standards, bridge organizational silos between different ministries and state governments, and establish a governance framework that operates with technological tools to establish and maintain public trust whilst accelerating AI innovation at scale.

Ultimately, DEPA for AI Training should be seen as more than a framework. It is India's next great contribution to the global commons, a second-generation Digital Public Infrastructure designed not just for transactions, but for trust. Building on the proven success of the India Stack, DEPA 2.0 offers a scalable, interoperable, and replicable model that can be adopted by nations and consortia worldwide. It provides the missing link that connects the principles of trustworthy AI to their practical, enforceable reality.