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

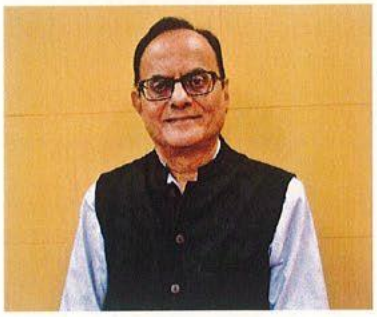
स्वच्छ भारत
एक कदम स्वच्छता की ओर

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

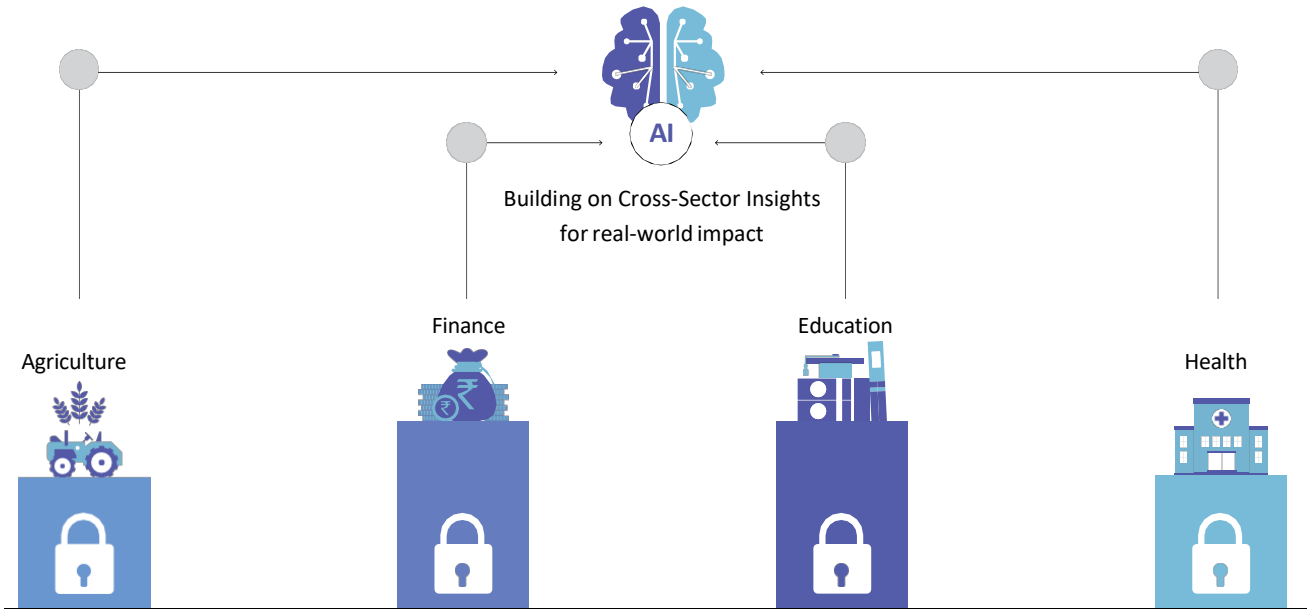
01 The Data Deadlock

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.

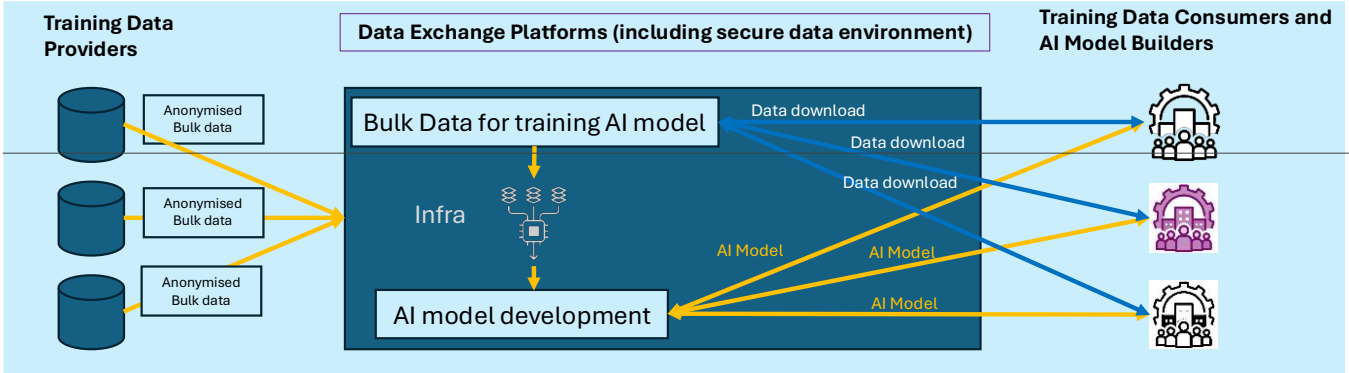
DEPA Training facilitates collaboration among key participants in the AI training lifecycle. Training Data Providers (TDPs) — institutions such as hospitals, banks, and public agencies holding valuable training data — collaborate with Training Data Consumers (TDCs) - encompassing innovators such as start-ups, 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.

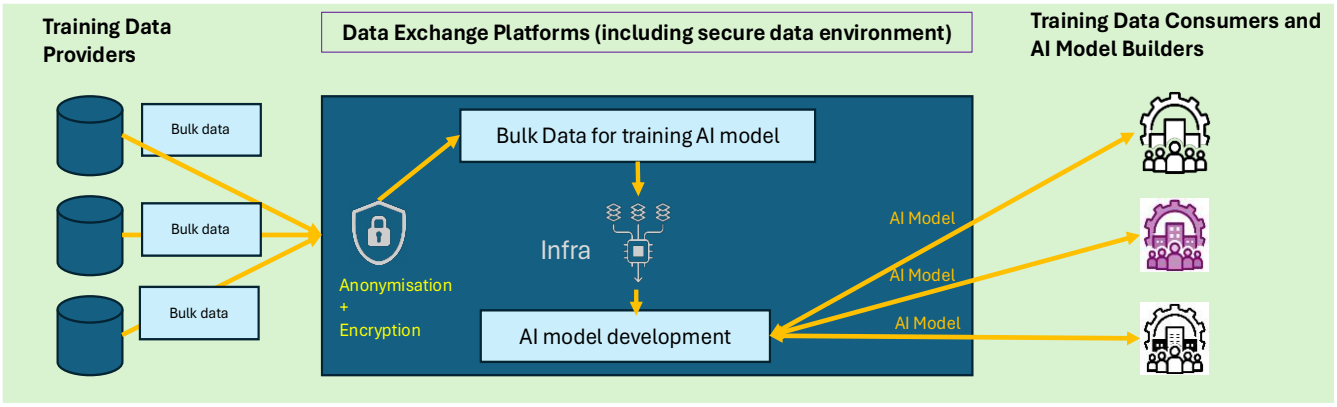
This framework provisions the use of Secure Data Environments along with Data Exchange Platforms for AI training, ensuring that TDPs' raw data remains protected in secure data environments and are unexposed to TDCs. Anonymisation of data is the responsibility entrusted to Training Data Providers. Training 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.

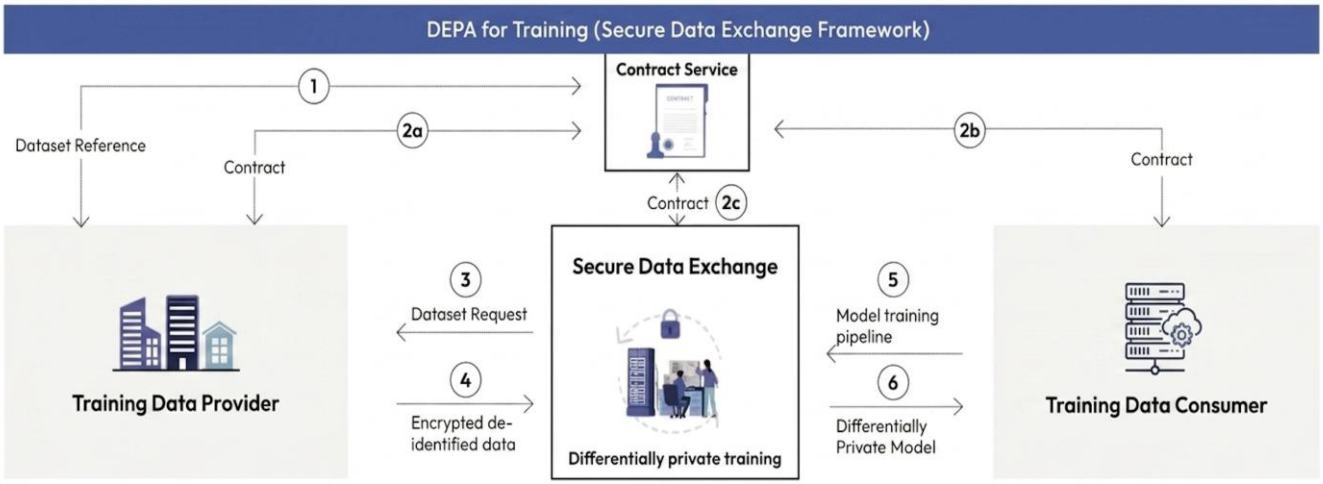
Option 1: Data Exchange Platform getting anonymized data from data providers



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

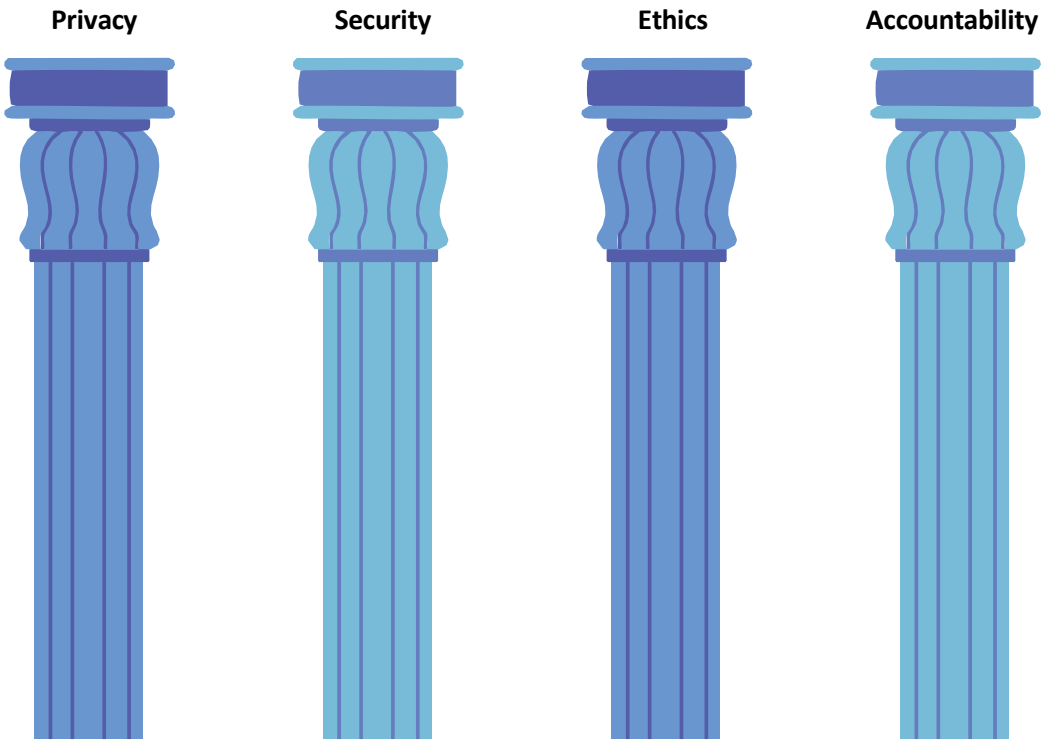
AI governance demands a fundamentally new approach—traditional static regulations simply cannot keep pace with this rapidly evolving, complex adaptive technology. India's solution leverages its proven expertise in building state-of-the-art Digital Public Infrastructures, adopting a techno-legal approach 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

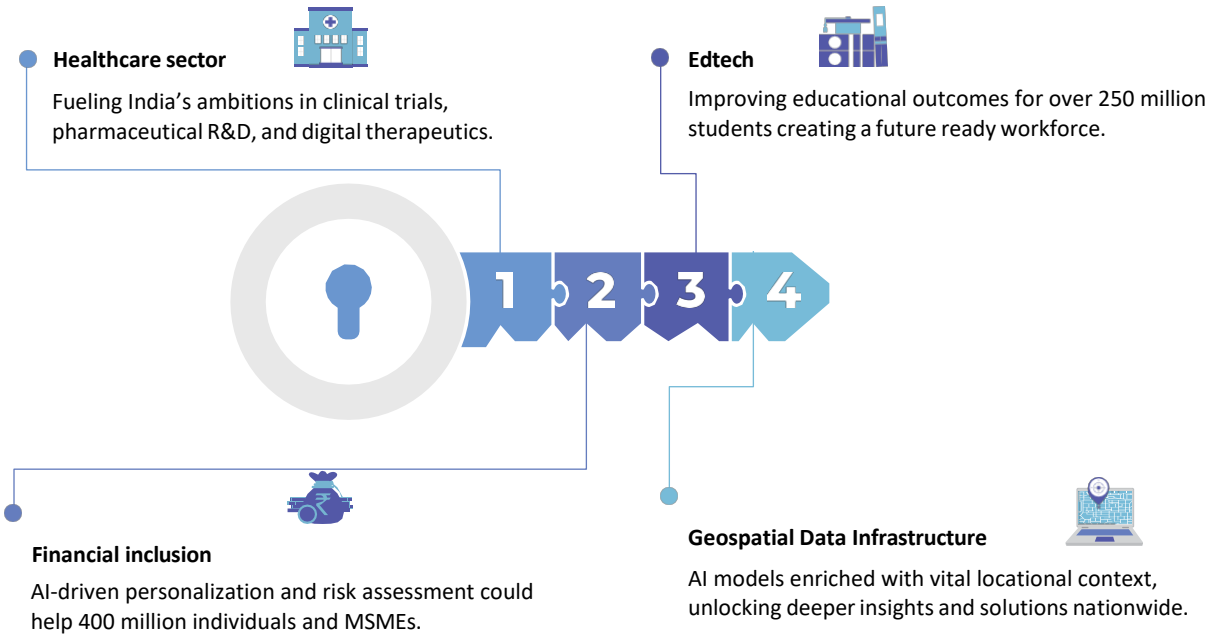
With DEPA Training architecture operationalized across sectors, India's DPI for AI stands poised to translate its unique digital advantages into leadership within the $15 trillion global AI economy. The framework provides the potential to unlock $500 billion in AI-driven value for India's GDP by 2030, 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.

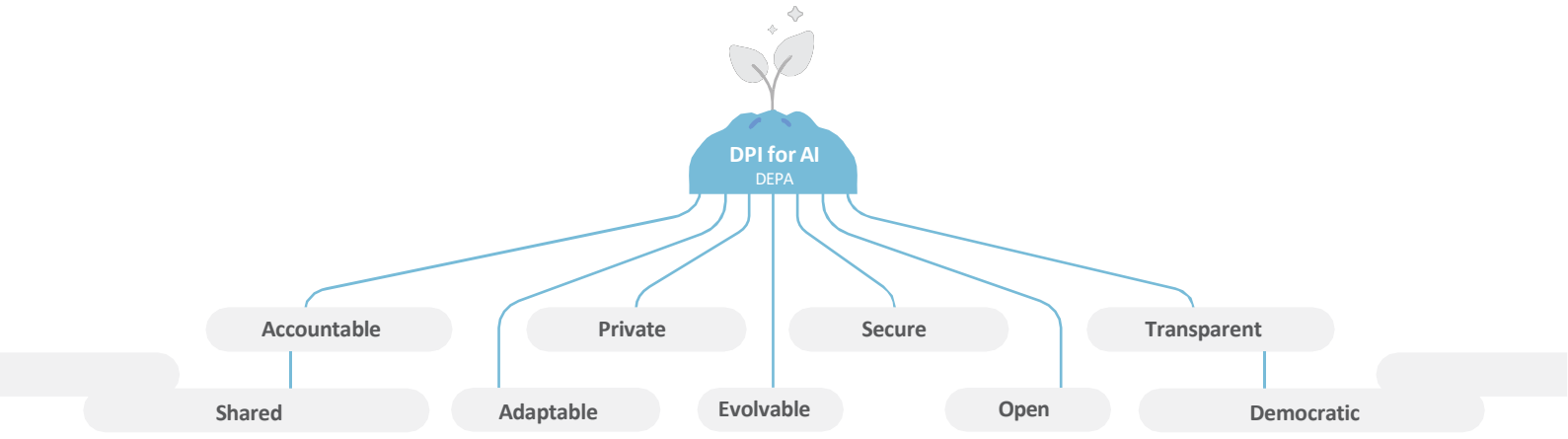


Table of Contents

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.

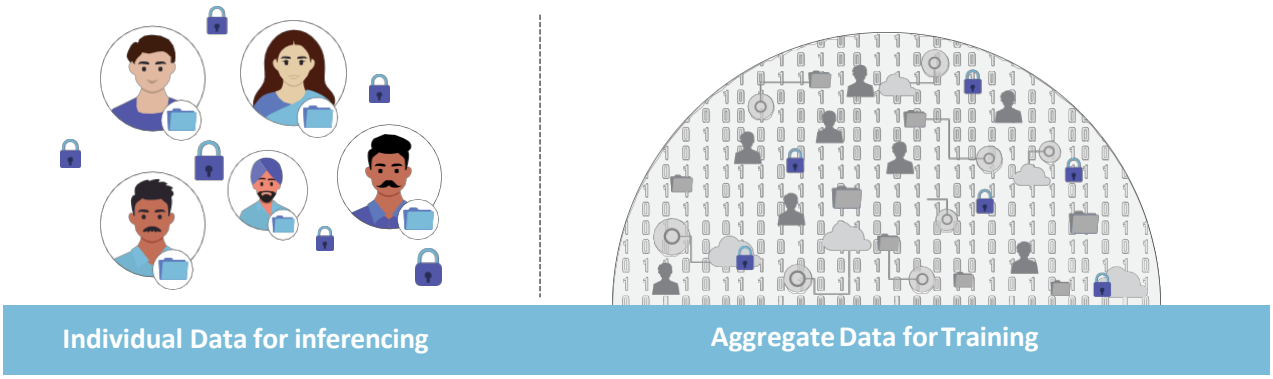
Importantly, a crucial characteristic of training data, especially for building sophisticated models like Large Language Models (LLMs) or complex predictive systems, is its **aggregate nature.** While it originates from individual data points (like a single transaction, a specific medical image, or a 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.

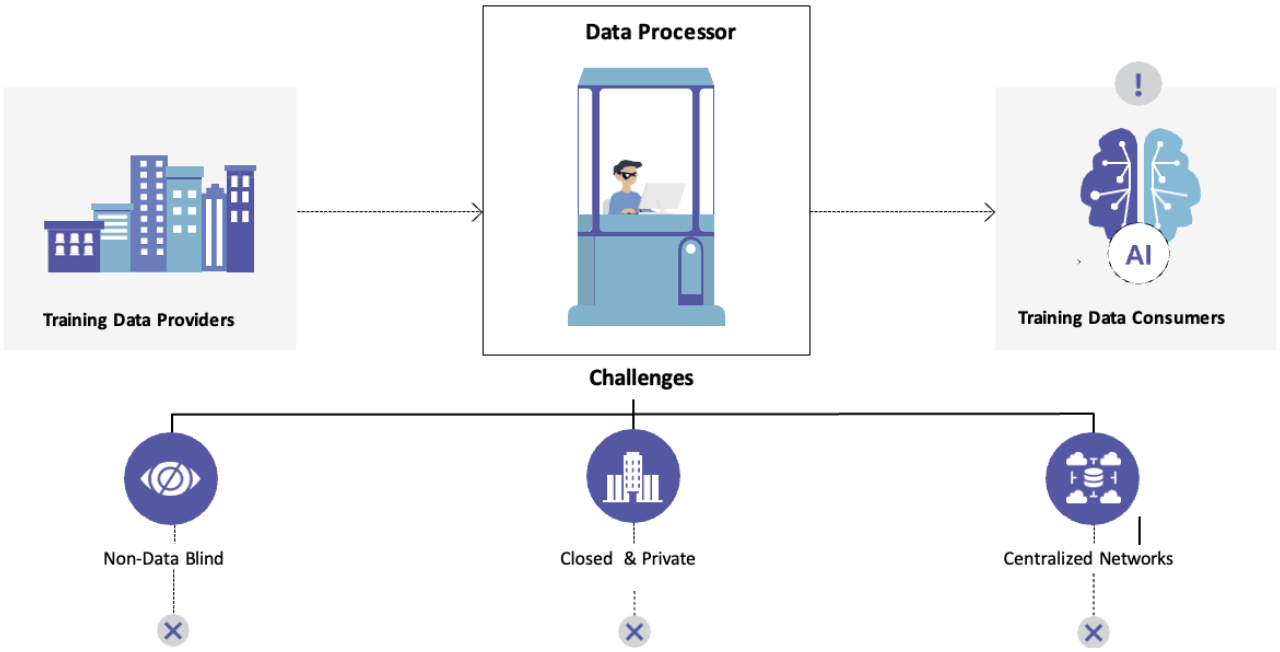
Third, the current situation could lead to **market concentration, and disadvantage local players.** Without a trusted and secure data collaboration framework there might be a create of an uneven playing field, where innovation becomes concentrated in the hands of a few. India possesses a key strategic 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.

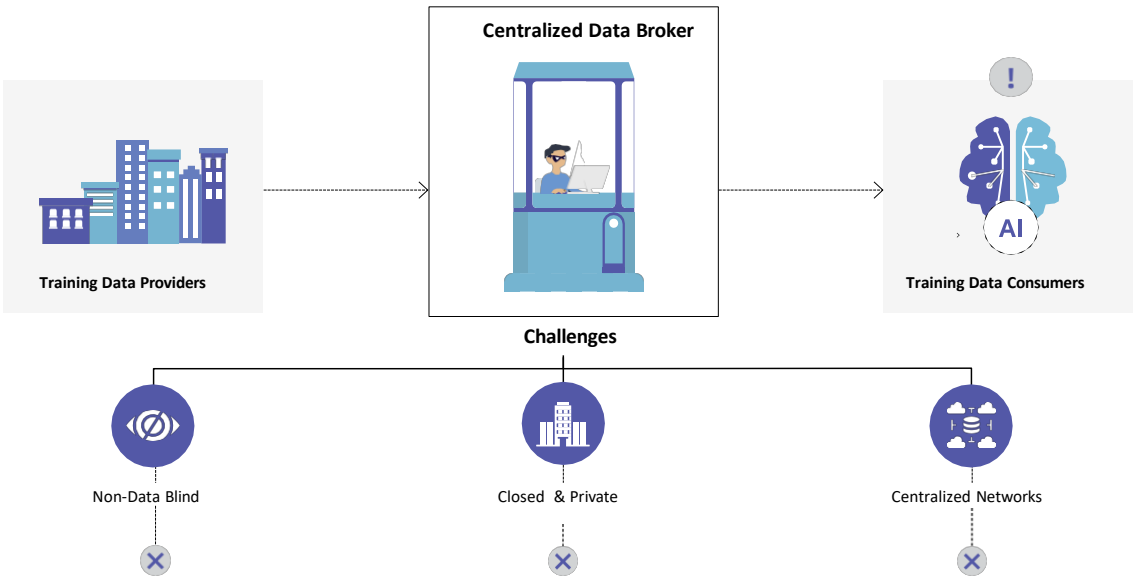
Compounding these structural issues are persistent **privacy vulnerabilities and unresolved intellectual property conflicts,** which collectively create a data deadlock, keeping valuable information locked in silos.

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

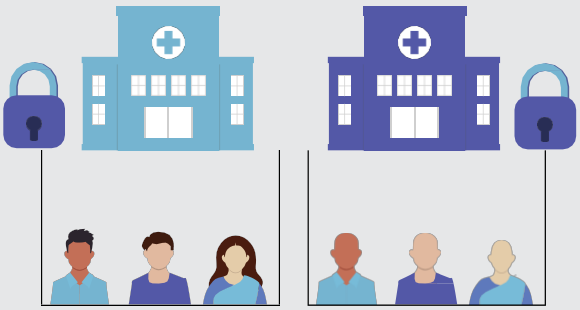
A common suggestion for overcoming data collaboration barriers is “anonymization” – simply removing direct identifiers like names, address- es, or unique ID numbers. The idea is that once PII is stripped, the data becomes non-personal and can be used for training freely without privacy risks. Unfortunately, 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

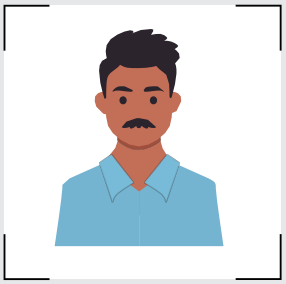


Valuable Health Data

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

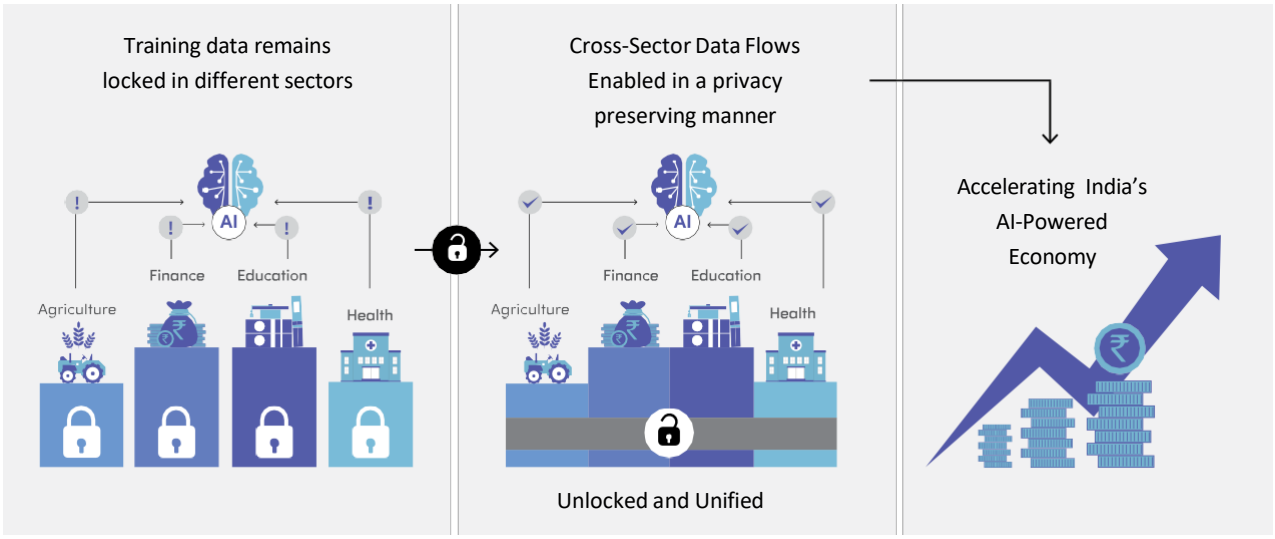
The global AI economy is expected to add over $15 trillion to global GDP by 2030, a transformation powered by data liquidity. India, with its billion-plus people online and an advanced stack of DPI is uniquely positioned to capture a significant share of this value. Projections indicate that India's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