



Enabling AI
development in
India through
data access

June 2024



Building a better
working world

Contents





04	●	Foreword
06	●	Executive summary
10	●	Background
16	●	Global AI leaders: approaches to data access
22	●	Data accessibility and availability: the India context
28	●	The way ahead for the public sector: recommendations
32	●	Recommendations for the private sector
36	●	Annexure 1: Data's role in bringing AI to life
37	●	Annexure 2: Indian data ecosystem

Foreword

AI advancement is driven not only by the sophistication of algorithms or access to computing power, but also by equal access to vast amounts of high-quality proprietary data. The speed of implementing newer use cases will depend on the availability of updated, relevant and customized datasets.

As countries and businesses work towards taking a leadership position in the development of AI, access to data will serve as a source of competitive advantage.

Access to data is not unique to India, and policymakers in major jurisdictions recognize its importance. AI development has been uneven across geographies, and the context of each country is different. Therefore, the approaches being followed vary. This report provides a high-level overview of the approach followed in other jurisdictions. It also analyzes the present scenario in the Indian context, including the initiatives taken by the Government of India, to provide access to both open data and proprietary data.

Until now, open-source and scraped datasets have provided a foundational framework for AI development. However, going forward, the development of AI use cases will require specificity and depth of data that comes through access to proprietary data. Lack of adequate and specific data for training AI algorithms could result in high error rates and bias, thereby impeding the development and uptake of AI. In an ideal scenario, the combination of the best AI algorithms, along with licensed access to high-quality and impactful datasets, can be a winning combination. Therefore, access to proprietary datasets will be a key differentiating factor between countries that emerge as winners and those that are unable to leverage the AI opportunity. This report makes recommendations to give effect to the above.





Those licensing data can benefit from monetizing their data and creating new revenue lines. The report discusses some of the steps that businesses can take to make sure that their data is protected, structured, properly collected and stored, and usable. Players with access to data would also need to ensure that privacy is protected and the process of both providing consent to process data and withdrawing consent is simple and easy to implement.

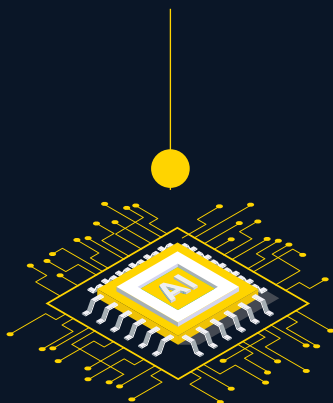
We hope that the rules and laws related to non-personal data in India will develop in a way that paves the way for open, fair and competitive markets. These, in turn, will drive innovation, create cutting-edge use cases, and underpin the future economic growth and success of India.



Rajnish Gupta

Partner, Tax and Economic Policy Group,
EY India

Executive summary



Data is at the heart of AI algorithms

Data drives the digital economy. Along with talent, computational and digital infrastructure, and algorithms, it forms the key building block of AI systems. The availability of the right type and amount of training data is a critical prerequisite for developing robust AI systems. As governments, businesses and researchers race to develop newer and more ambitious AI systems, the need to access larger and more specialized datasets is ever-increasing. This access is critical to meet the requirements of the use cases and the corresponding algorithms under development. Easier access to the right kind of datasets frees up time for researchers, allowing them to focus their attention on improving technology, creating new innovations and developing real-world applications.

Increasing availability of open data and incentivizing sharing of proprietary data aids AI development and innovation

Non-personal data can be available as open data or proprietary data. Open data is data that is available and accessible to the public at large for re-use. The availability of open datasets reduces the barriers to entry for AI development and innovation. Major jurisdictions around the world are taking measures to make more and more data available as open data to businesses, universities and individual researchers to spur development.

While businesses and governments think about data access, it is important to keep in view that data requirements and issues pertaining to data access will vary by industry and use cases. Further, some of this data may extend well beyond what is available as open data and may be available with private entities or specialized organizations. The wider ecosystem will not develop or innovate freely if access to data remains restricted.

Data whose title, ownership and control vests with one organization is referred to as proprietary data sets. These data sets may contain details about consumer behavior, product performance, financial information, and other specifics that can be leveraged to build AI algorithms. With the right safeguards and incentives, proprietary data can be used to obtain unique insights, build algorithms for targeted end uses, and ensure better compliance with privacy laws, regulations and security.

There are many impediments to sharing of data

Countries now recognize the importance of more data being made available for AI development. However, there are issues that need to be addressed to ensure easier access for a much wider usage:



Incentives for sharing of proprietary data



Ensuring privacy and confidentiality, especially in the case of personal data, even if it is anonymized



Clarity around ownership/ title of data holders, data intermediaries and the individuals



Ensuring that data management is not fragmented and there is interoperability between different AI systems/ data pipelines



Licensing and contracting frameworks that give confidence to individuals and small businesses that they are treated equitably by data collectors and intermediaries

Development of regulatory frameworks, platforms and institutional mechanisms will determine how data flows out from private repositories through either a licensing regime or a more open approach.

Governments are taking an active role of facilitating access to data

Recognizing the importance of making data accessible, several nations and governments are taking policy, regulatory and administrative measures.



Pillar I:

Institutional capacity and governance

- ▶ Establishment of a specialized agency for managing vast amounts of data and overseeing the development and management of the data ecosystem
- ▶ Investments to facilitate the development of ecosystem



Pillar II:

Privacy Framework

- Laws and regulations that ensure:
- ▶ Adequate privacy safeguards for personal data
 - ▶ Processing and use of personal data is with the consent of the user



Pillar III:

Facilitate access to non-personal proprietary data

- ▶ Frameworks/ regulations/ rules to promote sharing of proprietary data through marketplaces/ data exchanges, setting up interoperability standards, incentivizing private participation in data marketplaces and clarification on issues like title, etc



Pillar IV:

Making government data available

- ▶ Government has access to vast amounts of data which can be used as training datasets. Publishing data and digitizing these records as open data can enable the development of AI models in local languages

Executive summary

India has many opportunities to explore as it continues improving data access

The Government of India recognizes the importance of making data available widely. Most recently, the IndiaAI mission announcement in March 2024 underscores the significance of data in this mission. It goes on to state, "The IndiaAI Datasets Platform will streamline access to quality non-personal datasets for AI Innovation. A unified data platform will be developed to provide a one-stop solution for seamless access to non-personal datasets to Indian Startups and Researchers."

Over the past decade, the government has undertaken several initiatives in an effort to make data available.



The government operates an Open Government Data Platform that hosts data from various ministries and data collected through government surveys. Additionally, many Central and State government entities have made their data available through their own platforms. (ISRO through Bhuvan, NITI Aayog through NDAP, Ministry of Jal Shakti through India-WRIS, etc.)






India has enacted the Digital Personal Data Protection Act (2023). The Act provides clarity on the management of digital personal data, thereby both ensuring privacy of India's citizens and providing guidance to Indian businesses on how to deal with personal data.






India has also taken several individual steps such as 'account aggregator framework', "Bhashini", "DEPA", to name a few, to address data requirements in individual sectors/ departments.

For the future, here are some considerations the government could take into account:

-  Establishment of India Data Management, i.e., a central data regulator that could provide clear guidelines for data management and governance.
-  Maximize the availability of open data by undertaking a review of data presently available both at the Central and the State level, which could be leveraged for AI development. This could include examining the process for collecting and storage of data for future needs.
-  Identification, digitization and verification of official local language documents for training LLMs in regional languages.










1. <https://pib.gov.in/PressReleaseIframePage.aspx?PRID=2012355>

-  Accelerate the launch of data marketplaces - i.e., a centralized data marketplace for the exchange of data between both public and private entities following the norms already laid out by MeitY in its report "AI for India 2023". Incentivize data sharing among companies and individuals to promote access to a greater quantum of data, including financial incentives such as compensation for data shared.
-  Guidelines on the data title may help expedite the development of AI models by providing the private sector more autonomy and clarity on data usage. This could facilitate exchange of data on marketplaces, drawing from precedents in sectors such as the oil and gas sector where the title, rights and obligations of government and private players are clear.
-  Development of guidelines for data standards and interoperability will make data more accessible and usable for a wider audience.

Private players have opportunities to leverage value from access to data

Businesses usually collect data when they need it as part of their existing operations. Legacy efforts and systems are based on older practices or were designed for specific purposes. Businesses may have an opportunity to consider what all data can be generated, stored and leveraged for AI development. The development and access to datasets can be a catalyst for innovation and improving competitiveness.

While businesses may leverage data, there are concerns regarding data privacy and cybersecurity. In the event of cyberattacks or data breaches, there may be liabilities and reputational damages. Businesses have the opportunity to formulate a data-driven strategy to drive efficiency in their business operations by utilizing them in new AI tools. Some of the suggestions that can be considered are:

-  *Comprehensive data audit* could be undertaken at an organizational level. The audit would cover existing data practices, gaps, and potential data pipelines
-  *Generate more data* either internally or collecting external data through web scraping and other initiatives
-  Put in place the *process of editing, correcting and structuring data* that it is uniform in nature and prepared for analysis by machines
-  Enterprises must look to precisely *label and annotate their data* along with digitization, making their datasets AI-ready, helping algorithms recognize patterns and making predictions based on new, unseen data
-  Enterprises may automate a few of the manual functions from the wide range of approaches to data collection, processing, management, and retention
-  Since legacy IT systems have primarily been warehouse driven and data requirements for AI depends on both structured and unstructured data, a phased approach to transition to upgraded data systems and the approach to internal data governance should account for both legacy warehouses as well as the newer data
-  *Focus on digitizing all existing data in regional languages* which still remain in a physical form, including reports and forms, for machine learning, thereby enabling the development of AI tools that may also function in regional languages
-  *Data protection* should form an important part of any firm's data governance strategy wherein various tools like anonymization, ensuring consent, and auditing AI algorithms
-  Take measures to address *cybersecurity issues*



1

Background

Data is at the core of the digital economy and an essential building block for AI systems. It is one of the foundational pillars along with, talent, computational and digital infrastructure, and algorithms. The existing capabilities demonstrated by AI systems would not have been possible without access to large amounts of data.

The availability of the right type and amount of training data is a critical prerequisite for developing robust AI systems. As Governments, businesses and researchers set their sights on newer and more ambitious AI systems, there is an ever-increasing need to access larger datasets that meet higher quality threshold requirements for the different algorithms or the use-cases under development.

As more data becomes available, models can learn and adapt much better, thus helping improve the accuracy of AI generated content, while also enhancing the versatility

of AI systems (to handle varied requests). Large datasets have enabled improvement in the ability of AI systems to understand and process natural human language (speech recognition, sentiments and translation) and reduce biases, while improve decision-making ability of AI systems. The accuracy of AI systems is maintained by regularly feeding models with up-to-date information.

1.1 Availability of data has catalyzed the development of AI

The improved computation capabilities of AI systems have increased significantly with time. Consequently, data requirements have grown exponentially with the evolution and development of new age AI models such as large language models (LLMs). However, this was not always the case.

In the mid-2000s, the AI research ecosystem was mostly focused on AI models and algorithms. The ImageNet database project was one of the earliest instances of concerted efforts which laid specific emphasis on the development of an annotated dataset for training AI. ImageNet is a publicly available image dataset containing over 14 million labeled images. The ImageNet database served as a training dataset for neural networks in various computer vision tasks. Prior to the expansion and evolution of the ImageNet dataset to 14 million images, most datasets had approximately one million images.

Every year, during an annual ImageNet challenge, developers would train their own models using the ImageNet database. In 2012, AlexNet, a Convolutional Neural Network technology won the ImageNet 2012 challenge by achieving unprecedented accuracy, having reduced the error rate to 15.6% (post retraining). The success of AlexNet spawned a host of seminal applications of AI (leveraging Convolutional Neural Networks) relating to Image Detection, Object Detection, Semantic Segmentation, Facial Recognition, Art Generation, Medical Diagnosis (See Annexure-I for details).

The ImageNet Project was ground-breaking in demonstrating two key things:

Increase of data allowed for the development of newer and more accurate AI technologies

Enabling researchers to focus their attention on improving the technology, and developing new innovations and use cases, since the effort required to put together the data was already taken care of

Success of the ImageNet project demonstrates that data availability and accessibility are key to fostering innovation and facilitating the development of newer use cases.

1.2 Success of ImageNet database also illustrates the importance of open data

Non-personal data can be available as open data or proprietary data. Open data is data that is available and accessible to the public at large for re-use. On the other hand, proprietary data refers to that data for which the rights to title/ ownership are restricted, limiting the ability to freely access it.

Prior to the 2000s, the concept of open datasets was virtually absent. In the early 2000s, the Reuters-21578 text categorization collection was one of the first datasets to be freely available. The release of ImageNet in 2009 was also a significant event in the history of providing access to open datasets.

Since then, it has become increasingly common with large tech companies also contributing to open datasets (Tensorflow, PyTorch) and datasets such as Open Images and YouTube8m (Google), Facebook's AI datasets, Microsoft's MS Marco. Today, there are numerous open-source datasets that are available, such as Kaggle, UCI Machine Learning Repository, etc.

Availability of open datasets reduces the barriers to entry for AI development, and innovation.

The ImageNet project is a great example of how making one dataset available enabled a host of innovations as well as facilitated further development of AI.



1.3 The importance of proprietary data is growing

Proprietary data sets may contain details about consumer behavior, product performance, financial information, and other details that may be leveraged to build AI algorithms. Access to proprietary data provides the following benefits:

Unique insights: The accuracy and effectiveness of AI largely depends on the quality and quantity of the data it is trained on. Because this data is unique to the organization, it enables more personalized and precise predictions.

Better performance: AI models trained on proprietary data typically perform better for specific tasks tailored to the organization's needs compared to models trained on public data.

Innovation: Proprietary data can lead to the development of innovative AI applications specifically tailored to use cases. As an example, access to medical records can lead to the development of algorithms better equipped to predicting future health outcomes.

Privacy and control: By using proprietary data, organizations can maintain control over their AI training data. This can help ensure privacy, compliance, and security in the AI development process.

Continuous improvement: As more proprietary data is generated, AI algorithms can be retrained for improved performance.

In some cases, proprietary data can directly generate additional revenue for organizations. For example, a company can potentially sell or license its datasets or AI models (trained on its proprietary data) to other businesses or third parties. Consequently, there are marketplaces which have developed for some of these specialized/ annotated datasets.

Given the importance of data in AI development, there is a trend to maintaining and guarding proprietary data, as it will be valuable either through transactions in data marketplaces or by usage in the development of in-house algorithms and AI models.

1.4 The role of the government is key

Proprietary data sets act as a barrier to entry for newer organizations to enter the market or the AI ecosystem. These firms may be unable to compete with those with access to proprietary data sets to create AI models. Lack of mechanisms to make data widely available may lead to lack of innovation in some industries and impede competitiveness.

Therefore, governments play an important role in developing institutions and mechanisms that incentivize proprietary data sharing. To facilitate the same, development of data marketplaces may be a solution to enabling greater access to datasets.

Governments also have a role in opening up as much data as possible, so that the benefit may be derived widely. To this end, greater access to rich government datasets can also be leveraged for the development of India-specific AI use cases. Consequently, data platforms are being developed or considered for some of these specialized/annotated datasets.

The government's role is also key in devising frameworks pertaining to licensing of data, data rights, obligations of data intermediaries, etc. The government will also need to invest in the development of standards for interoperability of data to ensure data quality and completeness when it is being made available for the different use cases. All these measures contribute to the establishment of a vibrant data economy.

1.5 Data requirements vary based on the industry and use cases

While businesses and governments think about data access, it is important to keep in view that data requirements and issues pertaining to data access will vary by industry. The following table summarizes a few examples of AI use cases and their corresponding data requirements. These AI use cases include both new applications as well as the more traditional applications that may provide efficiency to current processes.

Table: Summary of various AI use cases across various industries and their specific data requirements

Industry	AI use cases	Data requirements
Financial services	Product and service design and innovation	Customer, market and competitor data, etc.
	Improving customer experience through Virtual Assistant enabled conversations	Summaries of past customer interactions, key concerns, FAQs, etc.
	Document creation for underwriting	Past documents, SOPs, product characteristics, etc.
	Marketing and sales	Email data, product documents, onboarding guides with text, audio and video, sales data, customer data, customer profiles, credit history etc.
	Collections, recovery and attrition control	User data, cashflow/bank statements, KYC data, payment schedules, regulatory requirements, risk management data, etc.
Healthcare	Clinical services and operations	Patient healthcare data and records, disease report databases, medical images, patient test reports, research data
	Community outreach	Data pertaining to key health concerns for AI-based content creation and personalized engagement, virus and disease data tracking for early detection of breakouts, etc.
	Audit and compliance	Regulatory data, digital forensics, compliance requirements etc.
Retail	Product, design and research	Product and packaging data, structures, material and ingredients data, design blueprints, etc.
	Procurement, manufacturing and quality assurance	Contracts data and documents, logistics and inventory data, quality control information, etc.
	Sales and marketing	Market research and product category data, product descriptions and insights, etc. User-level data to create buying recommendations, warranty, refunds and repairs data, etc.
	Store operations and staff management	Store level customer and inventory insights, product descriptions, etc.
Agriculture	Production planning	Soil mapping data, food supplies and prices data, subsidies data, provenance of crops, organic certification, other data points collected at wholesale markets for agriculture produce,, etc.
	Irrigation management	Groundwater availability, data about pipeline infrastructure and efficiency, data about power subsidies, historic rainfall information, soil moisture data, river flow data, flood data, soil moisture, etc.
	Crop protection and management	Data pertaining to crop insurance schemes, climatic forecasts, weather forecasts/pattern, crop storage infrastructure data, and data pertaining to costs and availability of pesticides and insecticides, historic disaster damage data, drone data, other data relevant for underwriting etc.



Table: Summary of various AI use cases across various industries and their specific data requirements

Industry	AI use cases	Data requirements
Technology services	Application development and support	Data on past coding models, automated response management, data engineering models, UI/UX designs, etc.
	Business process management	Processes data, customer experience data, etc.
	Infrastructure and operations	Contract data, past incident and response data, etc.
	Marketing	Marketing data, content data, customer data, etc.
Government services	Efficient policy drafting and data-driven decision making	Datasets on public consumption, expenditure, data on key governance indicators like education and healthcare, etc., government statistics, data report through M&E systems.
	Enhancement Citizen engagement	Personalized AI content, engagement on draft policies, government policies and schemes, grievance redressal, etc.
	Automated report generation	KPI tracking data, public expenditure and impact data, etc.
Media	Schedule and distribute content	Audio data for subtitles and captioning, rights management, fact checking for fake news, platform use data, etc.
	User engagement and monetization	User analytics/demographics, viewership tracking, user preference profiling/behaviors, etc.
	Prevent the spread of misinformation	Integrated databases to flag inaccurate news, social media streams, news streams, etc.

The requirements of data for AI development pertaining to the use cases extend well beyond what is available as open data. Some of the datasets pertaining to the use cases may be available with private entities or specialized organizations. The wider ecosystem will not develop or innovate freely if the access to data remains restricted. Thus, greater clarity would be required pertaining to incentive structures for sharing of propriety data, which will then determine how data flows out from private repositories through either a licensing regime or a more open approach.



1.6 Access to data is the key to AI: challenges to overcome

Countries now recognize the importance of more data being made available for AI development, especially given the great potential that AI has for economic impact. As a result, there are efforts being made to develop frameworks and regulations that enable access to a wider range of data for AI development, including non-personal data and anonymized personal data.

There is clear emphasis (both in India and across other countries) being made towards making more data available for AI developers, startups, and researchers. However, there are several challenges which need to be addressed to ensure that AI datasets are more easily accessible for wider usage.

Some of the key challenges are:

Incentivizing sharing of proprietary data: In the absence of financial and non-financial incentives, private businesses might not have a compelling reason to share data, especially since the data may be central to an entity's competitive advantage.

Ensuring privacy and confidentiality: Data may include information which could be personal and confidential in nature. Users may find it difficult to share such data if they are not sure who is going to use it, how it is going to be used and what sort of protections are in place.

Lack of clarity around ownership/ title and regulations: The rights and obligations of data holders, data intermediaries and the individuals need to be specified to simplify processes. The relevant regulatory and legal frameworks need to evolve to facilitate the same.

Fragmented data management and lack of interoperability between different AI systems/ data pipelines: Data may be inconsistent in terms of storage, management, structure and formatting. This may increase costs for gathering, cleaning and converting fragmented data into a more usable form. Currently, data for a particular use case may be available on several different systems. Interoperability of data between different AI systems is a key requirement. This becomes a particular challenge, especially when dealing with legacy systems with different proprietary database technologies.

Reservations towards licensing and contracting: Individuals and small businesses may have concerns that some contracts may be more favorable for data collectors and intermediaries.

Generation of large volumes of synthetic data: With the increase in the use of AI, there will be a lot of content that is generated from AI. It will be difficult to catalogue what is primary data and what is derived or generated through AI. Accordingly, the use of synthetic data may introduce other challenges such as biases, etc. As the volume of synthetic data available increases, systems will need to highlight if data is synthetic.

With this context, this whitepaper aims to provide recommendations for improving data access and quality of data available in the Indian domain so that it catalyzes AI development in India. The whitepaper is subsequently divided into four key sections as follows:

Global AI leaders: This section presents a global snapshot of what some of the leading countries are doing to make more data accessible. This allows for the creation of a framework to evaluate the Indian ecosystem.

Data accessibility and availability: the India context: This section covers key developments that have taken place in the Indian context, and the gaps that still remain.

Public sector recommendations: Finally, presenting the key measures that may be undertaken by the public sector to allow greater accessibility towards data with proper guidelines to overcome present challenges. This, in turn would allow India to become a leader in the AI and digital space.

Private sector recommendations: While the government will have a key role in providing frameworks and regulations for the AI ecosystem, the role of the private sector is key in making non-personal data available to ensure promote the development of the data ecosystem.





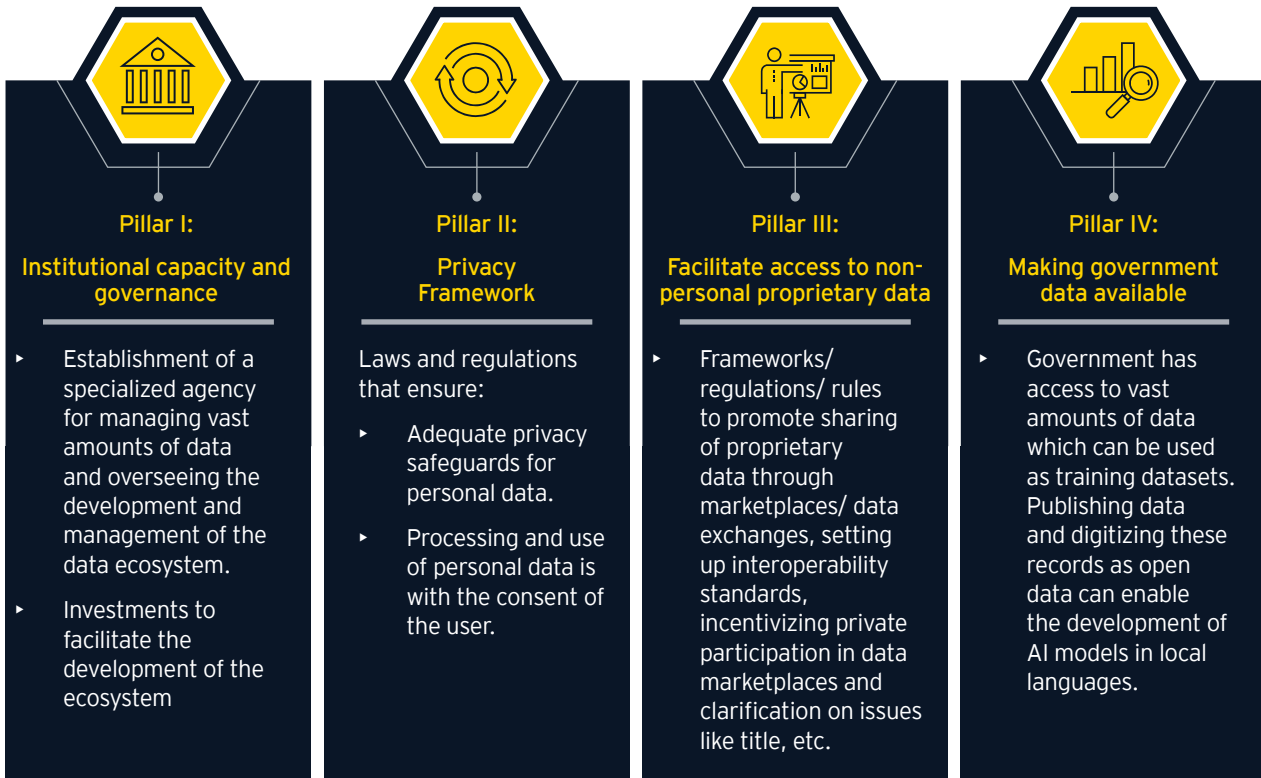
2

Global AI leaders: approaches to data access

With the growing importance of AI and the race to develop it to maintain competitive advantage, several nations and governments are taking measures to make data accessible to fuel AI innovation.

Our analysis suggests that there are four key thematic pillars which emerge relating to government initiatives towards improving data accessibility for AI development. These include, developing institutional capacity and governing the data ecosystem, establishing privacy frameworks, facilitating access to non-personal/private sector data, and making government data available. These are elaborated as follows:

the privacy of individuals. A series of measures are underway with the target to make larger amounts of data available through an institutionalized framework (through both re-use of publicly held data and the development of data intermediaries), improving voluntary data sharing in a trusted and safe manner, and providing greater control over non-personal data to generators of that data. The detailed regulations continue to evolve.



The approaches and frameworks developed by jurisdictions such as the European Union, United States of America and China are discussed in the subsequent sub-sections.



2.1 European Union (EU)

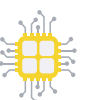
The EU recognizes the importance of making data more widely available, for economic and social benefit. It recognizes that data sharing is limited due to a number of obstacles and has publicly stated that 80% of high quality data is not used in the context of B2B and B2C data sharing. Accordingly, the EU has been developing a regulatory architecture with the objective of making the right data available for the right purpose, while protecting

2.1.1. Data Governance Act

EU's Data Governance Act entered into force on 23 June 2022 and has been applicable since September 2023 following a 15-month grace period. The act's provisions focus on enhancing trust in data sharing, improving data availability, and addressing technical barriers to data reuse. The act enables the creation of a framework that facilitates data sharing across sectors and EU countries while ensuring data privacy and security. The key highlights of the act as are follows:

Institutional mechanism: The act establishes the European Data Innovation Board to *“facilitate the sharing of best practices, in particular on data intermediation, data altruism and the use of public data that cannot be made available as open data, as well as on the prioritization of cross-sectoral interoperability standards.”*

3. European Union Final Report on Policy Conclusions, 2023



The European Data Innovation Board shall consist of at least the following three subgroups:

- ▶ **A subgroup is composed of the competent authorities for data intermediation services** and the competent authorities for the registration of data altruism organizations.
- ▶ **A subgroup for technical discussions** on standardization, portability and interoperability between data sets available platforms to improve usability and allow easier uptake of data on AI systems.
- ▶ **A subgroup for stakeholder involvement** composed of relevant representatives from industry, research, academia, civil society, standardization organizations, relevant common European data spaces and other relevant stakeholders and third parties advising the European Data Innovation Board on tasks which talks about maintaining records for data processing.

For access to open data, the EU has a European Data platform which presently provides access to 1.7 million datasets.

In conclusion, the institutional body may focus on providing comprehensive oversight for data processing and will cater to all key determinants of data accessibility, i.e., usage of proprietary data and public availability of open data.

Data intermediation services for facilitating access to proprietary data: The act provides for the setting up of data intermediaries which will operate as trusted organizers of data sharing or pooling within the common European data marketplaces. To enable the deployment of data intermediation services, the following provisions have been put in place:

- ▶ A set of rules for data intermediaries (neutral third parties) that connect individuals or companies with data users
- ▶ They can charge a fee for providing these services but cannot use the data themselves
- ▶ Data intermediaries must follow a set of rules to ensure neutrality and avoid conflict of interest
- ▶ Both stand-alone organizations and companies offering data intermediation in addition to other services could function as trusted intermediaries, and if data intermediation is a part of the offered services, the service provider needs to be legally and economically separated from other data services

Access to public data not available on open government platforms: The act requires member states of the European Union to be technically equipped to provide large volumes of data in a secure environment. The act discusses the steps that must be taken by a public body, if the access to any data cannot be granted publicly but may be required for AI usage.

It must ensure that privacy and confidentiality are protected by

- ▶ Using techniques such as anonymization, pseudonymization
- ▶ Accessing data in secure processing environments (e.g., data rooms)
- ▶ Confidentiality agreements
- ▶ Limit and avoid the use of exclusive data rights to specific cases of public interest
- ▶ Gives a public sector body two months to take decisions on reuse requests and authorities may also charge a reasonable fee for providing the data
- ▶ If data cannot be accessed, a public sector body should assist re-users find data by setting up a single point of information

2.1.2. Regulation on harmonized rule on fair access to and use of data

The European Data Act, which is formally known as the Regulation on Harmonized Rule on Fair Access To And Use Of Data, recognizes the transformative impact of data on various sectors of the economy. The act lays down the rules for accessing and using data in the European data economy. While the act was passed in January 2024, it will become applicable in September 2025 after an eighteen-month grace period. Most specifically, the European Data Act provides users with more control over their data and highlights the need for interoperability standards to be established in the data context.

- ▶ **User autonomy over generated data:** By giving users autonomy over generated data, the act addresses concerns related to data rights and obligations, data ownership and data accountability. Businesses and individuals are empowered to view, access, share, and modify generated data.
- ▶ **Interoperability standards:** Recognizing the importance of interoperability for fair data access and utilization, the act provides for the need for seamless interaction between different service providers. Specific Standards are expected to be formulated and released for the same going forward and the mechanics around the same are yet to be notified.

2.1.3. Regulatory control and safety (Global Data Protection Regulation)

The GDPR is a comprehensive European legislation which protects personal data/information. The GDPR was approved by the European Union in April 2016 and went into effect in May 2018. According to the act, "personal data may be understood as data which is identified to or can be identified to any person". Such data cannot be used/processed without appropriate anonymization or processes to convert it into non-personal data. The provisions of the act, which are enlisted below, regulate the way companies process and use personal data collected from consumers online.

GDPR applies to the entities in EU and pertains to usage of data related to the people of EU

- ▶ Objective is to safeguard data privacy by ensuring that EU subjects have complete security with regard to storage, processing, and transmission of their data
- ▶ The liability of data safety is placed on data users, who are required to restrict collection of personal data to the minimum requirement and ensure its deletion post usage
- ▶ GDPR also provides individuals the rights to access, amend, and erase their data



2.2 United States of America (USA)

United States has been the frontrunner in the development of AI globally. Much of the data used to train AI systems, particularly the development of large language models, has largely been scraped from the internet.

The focus of the government has been on "democratization" of access to data for individuals, researchers, universities, etc. through making greater amounts of data available publicly.

2.2.1 Data.Gov

Data.Gov is the United States government's open data website. which provides access to nearly 300,000 datasets which have been published by federal, state, and local governments along with contributions from universities and research organizations. Through the initiative, the government intends to enhance public accessibility to a wider array of databases, which may drive innovation, spur economic activity, and create greater transparency within the ecosystem.

2.2.2. National AI R&D Strategic Plan

The National AI R&D strategic plan outlines the United States strategy towards the development of fundamental and responsible AI systems. Given the need for data in AI development, the government intends to provide, fund and develop "Shared Public Datasets and Environments for AI Training and Testing". In the context of data availability for AI, the government intends to develop and make accessible datasets which cater to the diverse needs of AI development.



2.3 People's Republic of China

China has recognized the importance of access to data sets as an important building block for AI and has undertaken policy measures and administrative actions to improve access to data. In this regard, they have focused on developing institutional capacity, identifying impact sectors to ensure that data is available for developing algorithms in these sectors and taken steps to provide individuals with greater control over their data.

2.3.1. Institutional Capacity - National Data Administration

The National Data Administration was inaugurated in China in October 2023 under the purview of the National Development and Reform Commission. The institutional body is focused on regulating the utilization of data in China and is responsible for the coordination and building of a thriving data economy. It would also oversee the overall planning of the integrated sharing and development and use of data resources. As a part of these objectives, the National Data Administration has also developed a 'Data market focused three-year strategy called Data Element X Plan.

2.3.2 Data Element X Plan

Before the development and release of the Data Element X Plan, the National Data Administration carried out an exercise to identify priority sectors in which data integration can bring about a large-scale economic impact. 12 priority sectors, including industrial manufacturing, agriculture, commerce, transportation, financial services, technology, culture and tourism, healthcare, emergency management, weather services, urban governance, and green carbon, have been shortlisted after a thorough survey and audit of all sectors.



2.2.3. Personal Information Protection Law

The Personal Information Protection Law (PIPL) has been in effect since November 2021 and applies to both individuals and businesses. To address the protection of personal information and regulate data leakages, the law has the following key provisions:

- ▶ Individuals have the right to control their personal information, entities processing personal data must comply with legal requirements. This includes obtaining consent, ensuring data security, and providing transparency.
- ▶ Introduces personal information processors who will ensure that sharing of personal data is transparent and is with the due consent of the user

Recently, China has also started deliberating upon the need to establish systems for cross border data flows and standardize guidelines for the same.

Clearly, significant steps have been taken by the above-discussed countries to regulate and provide data for AI development. The laws and provisions highlight that the role of the government is crucial for the economies to take advantage of the AI revolution, since data is a fundamental building block of AI. In conclusion, with the advent and proliferation of AI, the role of governments with respect to access to datasets in the context of AI may broadly include the following roles/initiatives:

Key takeaways

01

The governments are developing ***institutional capacity*** to make more data accessible for use in developing AI systems. The institution may also be entrusted to establish a legal and regulatory framework to oversee the data ecosystem, which would enable easier trade of data.

02

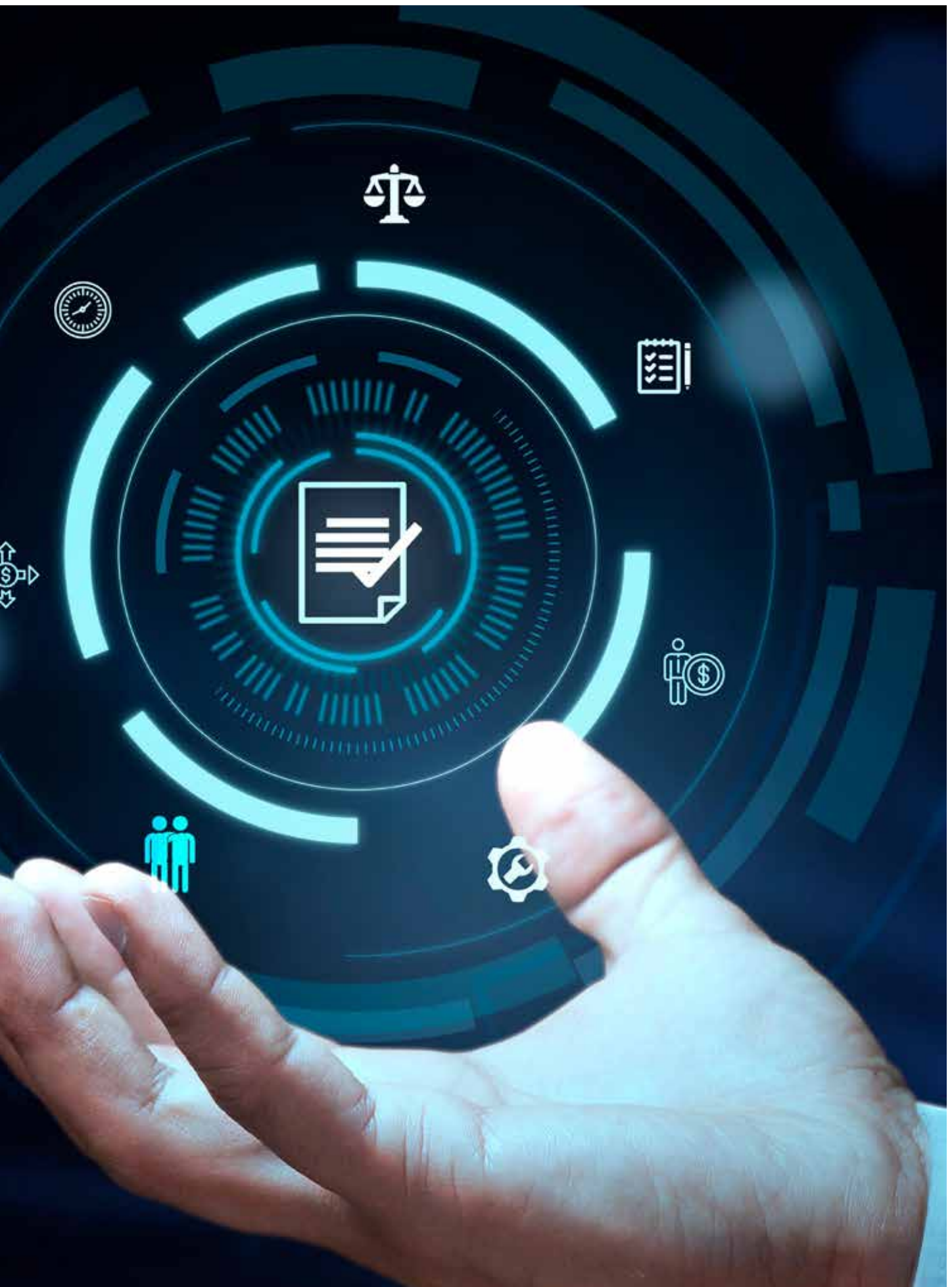
The governments have focused on facilitating proprietary data sharing to ensure that the ***right kind of data is available for the right purpose.***

03

The governments are endeavoring to ***publish a maximum amount of data available with them.*** Furthermore, public funding has been made for the development of datasets for AI applications, including but not restricted to AI development and testing.

04

The governments have either established or are establishing necessary ***guardrails for privacy and confidentiality*** of individuals and businesses while their data is used for AI development.





3

Data accessibility and availability: the India context

The various actions and plans of the Indian government towards data access, governance and management can be analyzed in the four-pillar framework, as discussed in the previous chapter. These four pillars include setting up institutional capacity, ensuring privacy, facilitating access to proprietary data through setting up of data marketplaces and maximizing the availability of public/ open data.

Table: Snapshot of various government initiatives related to access to datasets in India



Source: EY Analysis

The IndiaAI Mission has been established to facilitate the development and deployment of AI in India. The IndiaAI mission announcement in March 2024 made an allocation of INR10,300 crore towards the missions. Data is an important component of this mission, as it aims at creating the India Dataset Platform, a unified data platform to provide a one-stop solution for seamless access to non-personal datasets to Indian Startups and Researchers⁴.

Prior to the announcement in March, India recently enacted Digital Personal Data Protection Act (2023). The Act (discussed in Annexure) provides clarity on the management of digital personal data, thereby both ensuring the privacy of India's citizens and providing guidance to Indian businesses on how to deal with personal data. India has also taken several individual steps such as 'account aggregator framework', "Bhashini", "DEPA", to name a few, to address data requirements in individual sectors/departments, while few major initiatives like setting up of the India Data Management Office and India Dataset Platform are in the pipeline.

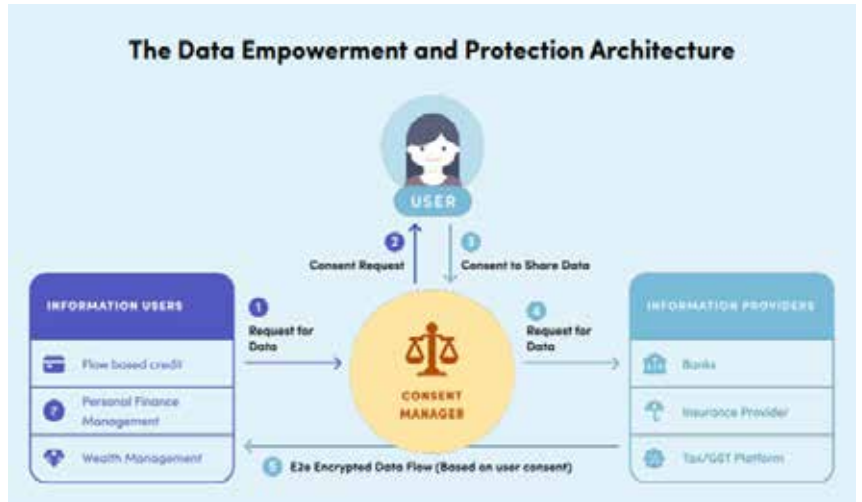
4. <https://pib.gov.in/PressReleaselframePage.aspx?PRID=2012355>



3.1 What India has done right: successful case studies

3.1.1. Citizen data and giving control of private data through consent: DEPA

The government introduced a seamless framework for processing personal information, which on one hand provides users greater authority towards the processing of their personal data and on the other hand makes provision for the data to be available for use in algorithms through both access and aggregation. To quote NITI Aayog, “despite increasing digitization, personal data (and particularly financial data) predominantly existed in fragmented silos”⁵, and DEPA aims to solve this issue.



Source: NITI Aayog

The **Data Empowerment and Protection Architecture (DEPA)** framework was introduced in 2020, wherein the concept of consent was brought in even before the Digital Personal Data Protection Act (DPDPA) 2023. This model provides an interplay between users, parties that hold user data such as banks and GST platforms, and parties which may process such data like personal finance and wealth management companies after procuring consent.

The DEPA framework defines how data would be routed through third party ‘Consent Managers’, thereby giving users greater control over their personal data. So far, this has been implemented in the financial sector under the joint leadership of the Ministry of Finance, RBI, PFRDA, IRDAI, and SEBI⁶.

Essentially, DEPA has helped create a framework that gives citizens control over their personal data and is also a key enabler for sharing personal data with third parties (account aggregators, data intermediaries, etc.). This framework can be leveraged in other sectors and may also help develop AI algorithms for these sectors.

3.1.2 Account aggregators as licensed intermediaries

The RBI has allowed account aggregators (AAs) to operate as licensed intermediaries that consolidate users’ financial data (customer/consumer). Account aggregation is a simple, fast and transparent way to promote data sharing between financial institutions that a user is seeking services/loans from, and with their consent, share with other institutions. AAs are data intermediaries who use Application Programming interface technology to collect, assemble, and synthesize data from multiple accounts in a single repository. It is a framework model (as discussed below) has been successfully implemented for the financial sector.



Source: Sahmati

5. <https://www.niti.gov.in/sites/default/files/2023-03/Data-Empowerment-and-Protection-Architecture-A-Secure-Consent-Based.pdf>

6. <https://www.orfonline.org/research/data-empowerment-and-protection-architecture-concept-and-assessment>

3.1.3 Bhashini: reducing digital divide across Indian languages

Bhashini was initiated by MeitY in 2022 to make technology accessible to all by removing language barriers. It aims to build a National Public Digital Platform for languages to develop services and products for citizens by leveraging artificial intelligence and other emerging technologies. Its platform brings together a large and diverse network including the government, industry, academia, research groups and startups to bring their contributions into an open repository of data⁷. The contributions shared are validated and standardized using a Unified Language Contribution API.

Bhashini is an important step to improve data access as it is creating a large open dataset by aggregating the contributions received by multiple stakeholders across different Indian languages into a shared repository. It is a much-welcomed initiative as machine learning and human translation are expected to work in unison to provide content in English and in one's native language to promote accessibility of digital services in India.

3.2 Key data governance initiatives under implementation in India

3.2.1 Institutional capacity: India Data Management Office

MeitY's Draft National Data Governance Framework Policy (2022) provides for the establishment of an India Data Management Office (IDMO) that would be responsible for framing, managing and periodically reviewing and revising the policy⁸. Presently, it is expected to be India's non-personal data regulator and comprises officials both from various ministries and experts from the private sector⁹. While it is still at a planning stage, the IDMO is expected to provide appropriate governance frameworks relating to: data integrity and audits, data regulation and the management of data. It would also provide anonymization rules and enable the creation of the India Dataset platform. These anonymization rules would be integral to making more data available by opening up of existing data that resides with both government and non-government entities.

7. MEITY: <https://www.meity.gov.in/writereaddata/files/Bhashini%20Whitepaper%20over%206.0.pdf>

8. MEITY: <https://www.meity.gov.in/writereaddata/files/National-Data-Governance-Framework-Policy.pdf>

9. Money Control: <https://www.moneycontrol.com/news/business/non-personal-data-regulator-idmo-to-have-both-govt-and-private-sector-representation-10794461.html>

10. IndiaAI 2023: MEITY

11. MEITY: <https://www.meity.gov.in/writereaddata/files/IndiaAI-Expert-Group-Report-First-Edition.pdf>

3.2.2 India's unified data sharing platform for non-personal data

India Dataset Platform (or IDP) (proposed initially in the National Strategy for Artificial Intelligence and then detailed out in the MeitY India AI Expert Group report of 2023) is envisaged to become a unified national data sharing exchange platform for all stakeholders to share and consume data/metadata/ artefacts/APIs without compromising their business or social goals, or on privacy, security and other concerns¹⁰. IDP is envisaged as an AI ecosystem which enables access to diverse high-quality datasets, encourage collaboration, help build data driven AI models, accelerate R&D and enable governments become data-driven organizations.

Some of the proposed features of the IDP include:

Data discovery

Single platform where data from different sources can be accessed and linked that includes both government and non-government sources¹¹.

Value-added services

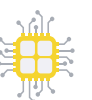
As a central agency, IDP will solve challenges faced by government departments and ministries at the initial state of data formation and help them curate their data while ensuring data quality and consistency.

Platform as a service model

Each domain group can independently manage and operate their data.

Federal structure

Different stakeholders will have autonomy and control over their data.



IDP plays dual role of improving data access and establishing data-governance

- 01 IDP lays the groundwork to operationalize a common data exchange platform. This will allow easier access of government and non-government data for AI developers, MSMEs, startups and academics.
- 02 The implementation priorities of the IDP include establishing governance, setting up data standards and formats, and developing security and privacy measures to enable scalability. The focus would be on regular monitoring for the improvement of the databases and the platform.

3.3 Legacy initiatives that tie into the data ecosystem

Over the past decade, government initiatives were focused on making government data available. Looking ahead, the underlying technologies and systems will require undergoing transformation to meet the demands and requisites of modern AI systems.

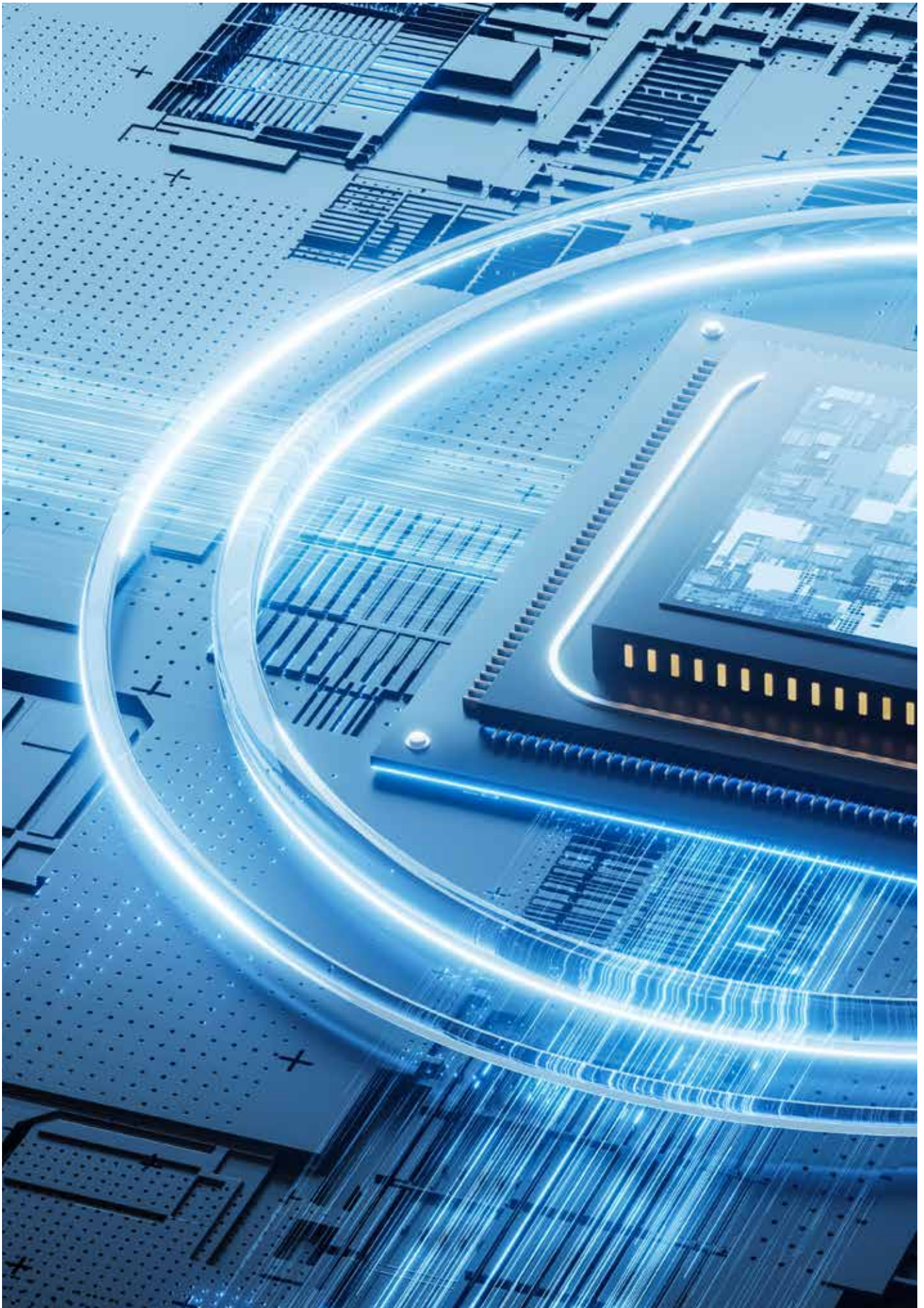
Open Government Data Platform: Presently, the government has an Open Data Policy and operates an Open Government Data Platform for India. The platform has data from various ministries and data collected through government surveys and published reports. There are numerous other Central and State government entities who have made their data available (ISRO through Bhuvan, NITI Aayog through NDAP, Ministry of Jal Shakti through India WRIS, etc.) A recent NASSCOM study sheds light on the issues faced by users in accessing this data and suggests that quality and comprehensiveness can be improved. To improve the usability of the data on this platform for AI development would require further investments.

12. NASSCOM: AI adoption index, 2022

Data available on public data platforms: Through the Digital India initiative, the government has focused on a digital first approach and has led to digitization of systems and processes across the entire government system. With the advent of IndiaStack, e-Office, DigiLocker, back-end upgradation etc. the government has been creating digital assets. Data available through these platforms could be leveraged, subject to legality, privacy, relevance, or specified use cases and other safeguards.

Healthcare data access: A good example of leveraging data on a public digital platform comes in the form of data generated through the implementation of the Ayushman Bharat Digital Mission (ABDM). The mission focuses on the delivery of healthcare services by integrating technology into hospitals nationwide. ABDM issues unique Health IDs to citizens, which serve as a digital identity for individuals within the healthcare system. The National Digital Health Eco-system thereby created supports the Universal Health Coverage in an efficient, accessible, inclusive, affordable, timely and safe manner through provision of a wide range of data and information. For the scheme to achieve its intended benefits, the digital public infrastructure, patient and medical data, and AI will be required to work in unison. Some of the critical use cases of the scheme are:

Use cases for AI in Ayushman Bharat	Various stages of scheme where data is integral
Internet of Medical Things	<ul style="list-style-type: none"> ▶ The scheme digitizes personal health records including prescriptions, diagnostic reports, medical histories, and billing information. ▶ It facilitates the seamless communication between healthcare intermediaries that will ensure that the records of the patients are portable.
Streamlining data, processes, and systems	<ul style="list-style-type: none"> ▶ Interoperability: AI ensures seamless data exchange between hospitals, simplifying processes and enhancing ease of living and treatment for patients. ▶ Digital consultations: Patients can access medical advice remotely through tech-powered digital consultations as records are readily available through digital means. ▶ Record management: All medical records are stored digitally, preventing loss and ensuring accessibility.





4

The way ahead for the public sector: recommendations

As discussed in the earlier chapter, there are several initiatives which help improve access to data. In addition, the government's efforts to regulate, manage and provide governance rules for data in India continue to evolve to create an enabling data ecosystem. There have been several working groups and draft papers that discuss approaches to various aspects of data access and governance in India. Given the Indian context, there are several challenges which need addressing, so that access to data can be improved:

Need for institutional oversight and regulatory clarity:



To overcome the new and intricate challenges that may arise with regards to data access and sharing, institutional mechanisms are necessary. The appropriate governance and regulatory mechanisms would also have to be evolved so that ambiguities pertaining to ownership of data and data standards/ interoperability can be addressed.

Keeping up with the global AI race and government data:



There is a global race to develop AI. Improved and faster access to specific data may help Indian startups and entities gain an edge. It can also enable the development of new and innovative India-specific

solutions. Thus, efforts to realize a functional and government backed data exchange/ platform/ marketplace for AI training data would be key to achieving the desired impact. The government's presence in this area is yet to be felt in the wider data ecosystem. Additionally, government-wide efforts to leverage and prepare data for AI training would not be getting the necessary push if there is no predetermined destination or process.

Translating data access to economic impact:



AI can have tremendous impact in some sectors, but the application of it may require sector-specific datasets for training AI models. Without the sector-specific data being available and without an in-depth understanding of the sectoral potential for economic impact through AI use, it may be difficult to achieve this. Furthermore, smaller companies/MSMEs may not have the resources or the technical expertise to develop and curate such datasets, leaving space for the government to intervene.

Proprietary interests and incentives to share:



Data collected by companies are often retained and not shared to retain a comparative advantage. Without the right incentives, financial or otherwise, getting private sector participation in data marketplaces and exchanges will be difficult with respect to participation.

4.1. Institutional capacity and governance

Establishment of the India Data Management Office (2023)

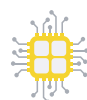
- Specifically, this requires a central data regulator in the form of an India data management office to take shape, which would provide clear guidelines on data management and governance for both public and private stakeholders.

Expedite the launch of data and AI marketplaces and data commons

- A centralized data marketplace or "India Dataset Platform" be created by India Data Management Office (IDMO) for exchange of data between both public and private entities following the norms already laid out by MeitY in its report "AI for India 2023".
- Beyond enabling access and availability of data, the marketplace would also have to bring clarity on certain aspects such as the terms of licensed use of datasets, whether reselling or further sharing of data be permitted, what disclosures would be necessary (with respect to purpose of use), and whether any limitations be applicable (eg. period of access) etc.

Publish guidelines towards data title

- "Data title" or data holding rights may be held with the government in some cases (like in terms of Oil and Gas exploration data), this is even if the data may have been collected by private entities.
- Guidelines on the data title for AI training datasets may help expedite development of AI models by giving the private sector more autonomy and clarity on the usage of data.



4.2 Access to proprietary non-personal data

Build data repositories for sectors with high impact use cases

- ▶ To help leverage AI, high-impact sectors may be identified, and industry wide data collection efforts may be considered.
- ▶ Various industries and sectors, such as healthcare, agriculture, logistics, etc. could contribute significant economic impact from the use of AI.

Incentives to share data through data marketplaces

- ▶ To promote access to greater quantum of data, companies and individuals need to be incentivized to share data. Incentives could be financial in nature such as consideration for data shared.

Standards and interoperability

- ▶ Development of guidelines for standards and interoperability of data will be crucial as it will allow the same data “accessible” and usable for a wider audience. It would for the allow use of data among different entities – both public and private, across different geographies and potentially even across sectors. There would also be compatibility for use across different AI systems.

Standardize and label public datasets

- ▶ Efforts may be undertaken to standardize, annotate and label public data as per applicability.
- ▶ The government may work with stakeholders to establish an overarching set of standards for data sharing and data maintenance. This will help ensure data completeness and interoperability between different AI model developers and cloud service providers.
- ▶ Simplify data pipelines, establish consensus mechanisms, and explore how labels can be audited to ensure accuracy/quality.

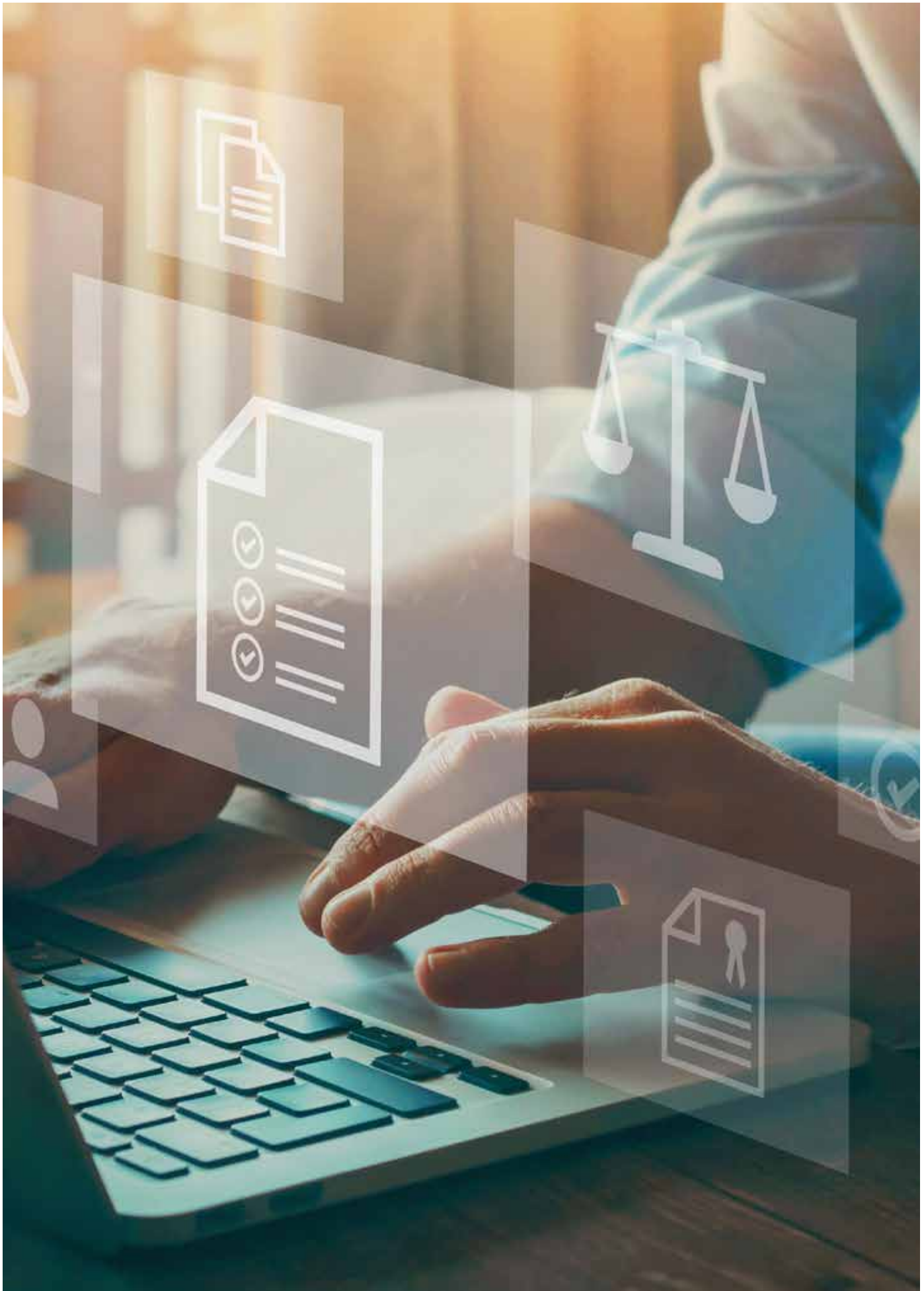
4.3 Making government data available

Consolidation of data available with the government

- ▶ Review of data presently available with the government that could be leveraged for AI development could be undertaken both at the Central and the State level. The review can also examine the process for collecting and storage of data for the future needs.

Digitization of historical official documents

- ▶ Increasing digitization has led to vast data generation in India, but language and cultural limitations and limited data availability in local languages restrict the development of localized Large Language Models (LLMs). To tackle this, a strategic plan involving identification, digitization, and verification of official local language documents for training AI is proposed. This may enable creation and improvement of region focused LLMs.





5

Recommendations for the private sector

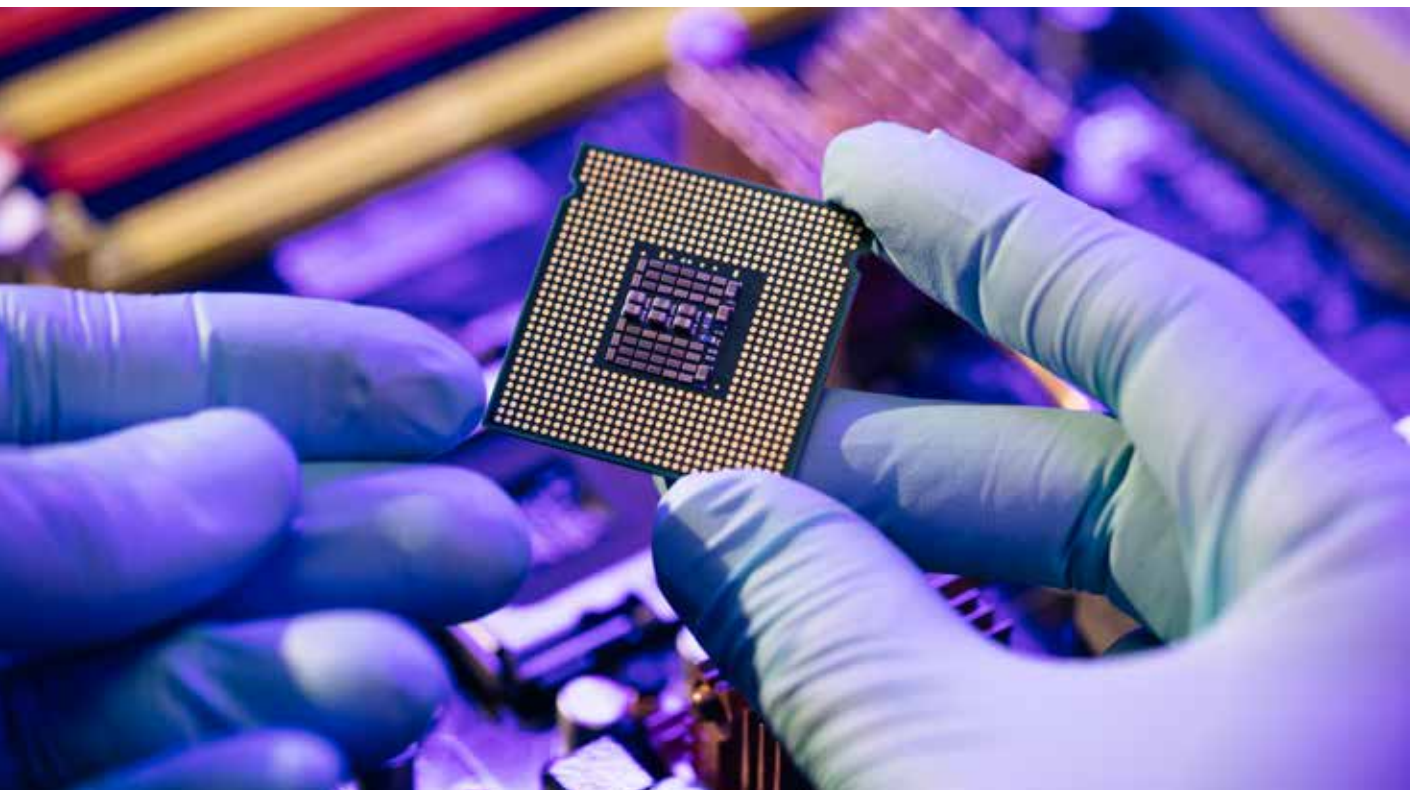
The countries which are leading in terms of AI have a strong private sector role. In some economies, the private sector has led innovation and development while the government has provided the legal and regulatory support. This highlights the importance of public-private-partnerships in the development of AI. The private sector has played a key role in the development of AI technologies. For instance, in the USA, AI development has primarily been promoted by the technology giants such as OpenAI, Microsoft, Google, etc. The private sector also plays an important role in terms of identifying and developing AI use cases. Accordingly, the private sector is often at the forefront of developing critical datasets, and in some instances have also made such datasets accessible for either the consumption of other businesses or for wider public use (as done by the Big tech companies).

However, for the Indian private sector at large, there are some broader challenges which remain:

🟡 **Need for streamlining data practices/ internal legacy challenges:** Businesses usually collect data when they need it as part of their existing business operations. These legacy efforts and systems which are in place either depend on older practices or were designed for specific purposes. They are not optimized and, in some instances, curated in a way that existing datasets are AI ready. Companies may also be lacking the technical infrastructure and capacity to be able to develop and curate AI ready datasets. These gaps may especially be daunting for businesses that who are relatively new to the prospect of developing and leveraging AI.

🟡 **Higher stakes, liabilities and risks around data:** The regulations around data creation, storage etc. are rapidly evolving especially for personal data and cross border data flows. It may be challenging for entities to keep abreast with such a rapidly evolving landscape, since it may require continuous adherence. Additionally, in the event of cyberattacks or data breaches, there may be liabilities and irreparable damages to the business which must be considered. These gaps continue to remain unaddressed.

🟡 **Loss of potential revenue and opportunities:** Proprietary private sector datasets are often not shared for competitive reasons. As a result, companies potentially lose out on revenue that may have been obtained through licensing and participation in data marketplaces. This also limits the potential for the development of new innovations and collaborations. The development and access of datasets is often a catalyst for innovation. These benefits are limited in a closed data ecosystem.



The following recommendations aim to address the broader issues in the context of the private sector and access to data for AI development:

5.1 Standardization of datasets

Collection, cleaning and standardization of datasets

- ▶ Generate more data either internally or collecting external data through web scraping and other initiatives
- ▶ Put in place the process of editing, correcting and structuring data that it is uniform in nature and prepared for analysis by machines
- ▶ Remove corrupt or irrelevant data and formatting it into a language that computers can understand
- ▶ Formulate a data driven strategy to drive efficiency in business operations, through utilizing the same in new AI tools developed to solve business related issues

Labeling and annotation

- ▶ Labeling and annotation helps algorithms recognize patterns and make predictions based on new, unseen data
- ▶ Enterprises must look to precisely label and annotate their data along with digitization, making their datasets AI-ready.
- ▶ Some of the best practices for data labeling have been mentioned in the Annexure

Automation and data pipelines

- ▶ Enterprises have had a wide range of approaches to data collection, processing, management and retention, some of which may have been manual and need to be automated for data to be optimized for AI use
- ▶ Automating the data pipelines will help organizations derive the greatest benefit of AI

5.2 Securing existing datasets

Unified data governance

- ▶ Data requirements for AI depend on both structured and unstructured data. The legacy IT systems have primarily been warehousing driven. The approach to internal data governance should account for both legacy warehouses as well as the newer data lakes

Ensuring protection of private datasets

- ▶ Data protection should form an important part of any firm's data governance strategy, wherein various tools like anonymization, ensuring consent, and auditing AI algorithms may be used
- ▶ Firm's data guidelines to conform to DPDPA 2023

5.3 Enabling access to datasets

Digitizing datasets, including those in regional languages

- ▶ Digitize all existing data which still remains in a physical form, including reports, forms, etc. for machine learning
- ▶ Focus should be digitizing the data which exists in regional languages, thereby enabling the development of AI tools that may function in regional languages as well

Handling of legacy systems

- ▶ Organizations may vary greatly in terms of the legacy systems and a phased approach may be adopted to transition to upgraded data systems

Participation in data marketplaces

- ▶ Firms may bring forth their anonymized, well-structured and labeled datasets to central data marketplaces that are being developed by the government
- ▶ Derive value out of data provided to the marketplace and purchase data relevant to the firm's strategy and requirement
- ▶ Ensure privacy of users is intact and that the India government's data governance strategy is followed

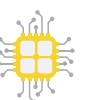
5.4 Review and compliance measures for existing datasets

AI data audit and strategy

- ▶ A comprehensive data audit could be undertaken at an organizational level to develop an AI focused data strategy
- ▶ The audit could cover existing data practices, gaps and potential data pipelines that need to be established

Review and compliance

- ▶ Ensure that the developments with respect to compliance requirements and interoperability are regularly monitored and adhered to
- ▶ Establish systems and SOPs that will ensure that compliance requirements are factored in designing internal organization processes



Annexure 1:

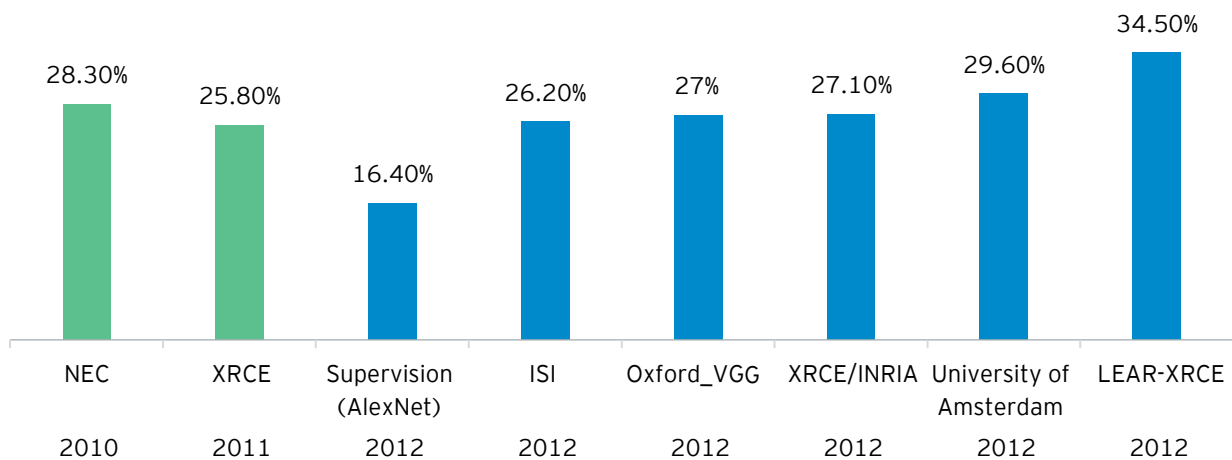
Data's role in bringing AI to life

Table: Seminal use cases that have emerged from convolutional neural networks (CNNs)

Application	Definition	Examples
Image detection	In image classification, CNNs are used to categorize images into predefined classes or labels	Identifying objects in photos for social media platforms, detecting spam or inappropriate content in image-based email
Object detection	Object detection involves identifying and localizing multiple objects within an image	Self-driving cars use CNNs to detect pedestrians, vehicles, and traffic signs to make real-time decisions. Retail stores employ object detection to track inventory levels and prevent theft
Semantic segmentation	Semantic segmentation assigns a label to each pixel in an image, creating a detailed segmentation map.	Autonomous drones use semantic segmentation to navigate obstacles and avoid collisions
Face recognition	Face recognition systems identify and verify individuals based on facial features	Face unlocks in smartphones and security systems
Style transfer and art generation	CNNs can transform images by applying artistic styles to them.	Personalized artwork and recreating digital art
Medical diagnosis	CNNs analyze medical images (X-rays, MRIs, etc.) to assist doctors in diagnosis	Medical imaging, where CNNs segment tumors or organs from MRI or CT scans

AlexNet: winner of ImageNet challenge

By acing the challenge, AlexNet was able to highlight the practical aspects of deep learning and increase the trust of the users in the same. The success of AlexNet would not have been possible without the dataset being made available by ImageNet. While most technologies performed well on smaller databases, the error rate increased drastically when the same technologies were tested on larger datasets. The increased accuracy achieved by AlexNet laid the groundwork for several uses of AI based image recognition that have since gained popularity and have had a profound impact on society.



Graph: Error rates of challenge winners from 2010, 2011 and the top six results from the 2012 challenge

Note: Post retraining, the accuracy rate of AlexNet reduced further to 15.3%

■ Trained on 1 million images
 ■ Trained on 14 million images

Annexure 2:

Indian data ecosystem

1. Digital Personal Data Protection Act (DPDPA) 2023

DPDP Bill 2023 was passed by the Indian parliament in 2023, and subsequently became an Act in August 2023, laying the guidelines for processing personal data while safeguarding the privacy of digital personal data being generated by the Indian populace. Essentially, the DPDPA 2023 recognizes the right to individual privacy in the following three ways:

- ▶ Clearly entailing the obligations of data fiduciaries who are the entities who process data
- ▶ Defining the rights and duties of Data Principals, who generate the digital data
- ▶ Demarcating the financial penalties for breach of rights, duties and obligations

What this Act brings forth in a clear way is to demarcate the necessity of consent to be taken by data fiduciaries, the exceptions wherein personal data may be processed by the government and also introducing the Data Protection Board of India. Regardless of whether the data principle is able to carry out their own duties to protect their personal data, the data fiduciaries are being held responsible in a big way to ensure right to privacy.

2. Data governance initiatives that are yet to be implemented in India

A. Draft National Data Governance Framework Policy (2022)

The Draft National Data Governance Framework Policy was launched in December 2022 to provide an overarching framework for India's approach to data governance.

Objectives	Applicability	Initiatives
<ul style="list-style-type: none"> ▶ Digital governance ▶ Standardized data management ▶ Promoting transparency, accountability and ownership of non-personal data ▶ To set quality standards ▶ Expand India datasets program 	<ul style="list-style-type: none"> ▶ All government departments under central government ▶ All non-personal datasets access and use by researchers and Start-ups ▶ State governments to be also encouraged 	<ul style="list-style-type: none"> ▶ Setting up of India Data Management Office (IDMO) ▶ IDMO shall formulate all data/datasets/ metadata rules, standards, and guidelines in consultation with various ministries and the industry ▶ IDMO will process requests and provide access to non-personal data to researchers and startups, thereby encouraging AI based research

B. Draft India Data Accessibility and Use Policy (2022)

The India Data Accessibility and Use Policy published as a draft in 2022, aims at enhancing the access, quality and use of data, to meet the emerging technology requirements of the decade. The key features of the policy, including the proposed institutional framework and guidelines, are the following:

- ▶ India data office (IDO) to be set up, which would streamline and consolidate data access and sharing of public data repositories across government and private stakeholders. IDO would also identify and allow access to high-value datasets.
- ▶ All government ministries would adopt the high value dataset (HVD) framework to identify, publish and maintain these datasets which have a high degree of importance in the market and have socio-economic significance.
- ▶ For restricted access to data sharing, the pricing would be determined by the concerned government department.
- ▶ Data standards would be determined by the India Data Council, and these would be adopted by all government departments and ministries. Minimum anonymization standards would need to be complied with.



- ▶ Data retention period for specific datasets would need to be determined and complied with by all government departments and ministries.

3. Data interoperability and standards for the AI ecosystem

Interoperability in AI datasets facilitates data integration between different systems which have access to similar data. This ensures that data availability is not fragmented. The government released the Draft Interoperability Framework for E-Governance in October 2015. The framework focuses on describing an approach to overcome challenges that the government faces in enabling joint delivery of public services.

BIS has introduced a standard for AI (IS/ISO/IEC TS 4213: 2022 Information Technology – Artificial Intelligence – Assessment of Machine Learning Classification Performance). This pertains to “measuring classification performance of machine learning models, systems and algorithms.” Similar efforts can be replicated to establish standards for datasets to enhance their usability across platforms.

AI standards are globally being developed through ISO/IEC and other organizations. As the technology and its application evolve, so will the standards. The present standard adopted by BIS is focused on the classification of “performance” of ML systems. Subsequently, issuing standards around data will be useful to both government and private entities as they develop interoperable future-ready datasets.

Acknowledgments

Steering Committee:

Mahesh Makhija
Rakesh Kaul Punjabi
Rajnish Gupta
Vineet Mehta
Alexy Thomas

Core Team:

Rajnish Gupta
Ankan De
Shambhavi Sharan
Rishi Dewan

Editorial Team:

Prosenjit Datta
Vikram D Choudhury

Design Team:

Ridhi Sharma Kapuria

Our Offices

Ahmedabad

22nd Floor, B Wing, Privilon
Ambli BRT Road, Behind Iskcon Temple
Off SG Highway
Ahmedabad - 380 059
Tel: + 91 79 6608 3800

Bengaluru

12th & 13th Floor
"UB City", Canberra Block
No.24 Vittal Mallya Road
Bengaluru - 560 001
Tel: + 91 80 6727 5000

Ground & 1st Floor
11, 'A' wing
Divyasree Chambers
Langford Town
Bengaluru - 560 025
Tel: + 91 80 6727 5000

Bhubaneswar

8th Floor, O-Hub, Tower A
Chandaka SEZ, Bhubaneswar
Odisha - 751024
Tel: + 91 674 274 4490

Chandigarh

Elante offices, Unit No. B-613 & 614
6th Floor, Plot No- 178-178A
Industrial & Business Park, Phase-I
Chandigarh - 160 002
Tel: + 91 172 6717800

Chennai

6th & 7th Floor, A Block,
Tidel Park, No.4, Rajiv Gandhi Salai
Taramani, Chennai - 600 113
Tel: + 91 44 6654 8100

Delhi NCR

Ground Floor
67, Institutional Area
Sector 44, Gurugram - 122 003
Haryana
Tel: +91 124 443 4000

3rd & 6th Floor, Worldmark-1
IGI Airport Hospitality District
Aerocity, New Delhi - 110 037
Tel: + 91 11 4731 8000

4th & 5th Floor, Plot No 2B
Tower 2, Sector 126
Gautam Budh Nagar, U.P.
Noida - 201 304
Tel: + 91 120 671 7000

Hyderabad

THE SKYVIEW 10
18th Floor, "SOUTH LOBBY"
Survey No 83/1, Raidurgam
Hyderabad - 500 032
Tel: + 91 40 6736 2000

Jaipur

9th floor, Jewel of India
Horizon Tower, JLN Marg
Opp Jaipur Stock Exchange
Jaipur, Rajasthan - 302018

Kochi

9th Floor, ABAD Nucleus
NH-49, Maradu PO
Kochi - 682 304
Tel: + 91 484 433 4000

Kolkata

22 Camac Street
3rd Floor, Block 'C'
Kolkata - 700 016
Tel: + 91 33 6615 3400

Mumbai

14th Floor, The Ruby
29 Senapati Bapat Marg
Dadar (W), Mumbai - 400 028
Tel: + 91 22 6192 0000

5th Floor, Block B-2
Nirlon Knowledge Park
Off. Western Express Highway
Goregaon (E)
Mumbai - 400 063
Tel: + 91 22 6192 0000

3rd Floor, Unit No 301
Building No. 1
MindSPACE Airoli West (Gigaplex)
Located at Plot No. IT-5
MIDC Knowledge Corridor
Airoli (West)
Navi Mumbai - 400708
Tel: + 91 22 6192 0003

Pune

C-401, 4th Floor
Panchshil Tech Park, Yerwada
(Near Don Bosco School)
Pune - 411 006
Tel: + 91 20 4912 6000

10th Floor, Smartworks
M-Agile, Pan Card Club Road
Baner, Taluka Haveli
Pune - 411 045
Tel: + 91 20 4912 6800

Ernst & Young LLP

EY | Building a better working world

EY exists to build a better working world, helping to create long-term value for clients, people and society and build trust in the capital markets.

Enabled by data and technology, diverse EY teams in over 150 countries provide trust through assurance and help clients grow, transform and operate.

Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. EYG member firms do not practice law where prohibited by local laws. For more information about our organization, please visit ey.com.

Ernst & Young LLP is one of the Indian client serving member firms of EYGM Limited. For more information about our organization, please visit www.ey.com/en_in.

Ernst & Young LLP is a Limited Liability Partnership, registered under the Limited Liability Partnership Act, 2008 in India, having its registered office at Ground Floor, Plot No. 67, Institutional Area, Sector - 44, Gurugram, Haryana - 122 003, India.

©2024 Ernst & Young LLP. Published in India.
All Rights Reserved.


EYIN2406-003

ED None

This publication contains information in summary form and is therefore intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. Neither EYGM Limited nor any other member of the global Ernst & Young organization can accept any responsibility for loss occasioned to any person acting or refraining from action as a result of any material in this publication. On any specific matter, reference should be made to the appropriate advisor.


RS1


ey.com/en_in

 @EY_India

 EY

 EY India

 EY Careers India

 @ey_indiacareers