TAKSHASHILA INSTITUTION

A Pathway to AI Governance

Bharath Reddy
Nitin Pai
Rijesh Panicker
Satya S. Sahu
Sridhar Krishna

This discussion document critically examines the different stages of the AI supply chain exploring a pathway for AI governance from a national interest perspective.

Takshashila Discussion Document No. 2024-10
Version 2.0, June 2024

Recommended Citation:
Bharath Reddy, Nitin Pai, Rijesh Panicker, Satya Sahu, Sridhar Krishna, " **A Pathway to AI Governance**," Takshashila Discussion Document No. 2024-10, June 2024, The Takshashila Institution.

Executive Summary

Artificial intelligence has immense potential to enhance human capabilities and drive growth in several industries. It can also greatly improve education, healthcare, and governance outcomes, particularly benefiting low-income countries. However, this potential may not be realised if the AI market remains concentrated in the hands of a few dominant players.

While global AI governance efforts primarily focus on ethics and safety, it is crucial to consider AI governance from a national interest perspective. This involves examining how AI adoption can help humans flourish, strengthen democracy, and promote a stable global order. It also looks at the need for sustainable practices in AI adoption and the importance of ensuring competition in the AI ecosystem. These considerations need to be examined across these different stages of the AI supply chain - data, computation, model, and application - to envision the desired outcomes at each stage.

At the data stage, a marketplace that recognises individual ownership of data and empowers individuals to dictate the usage and distribution of their personal data is essential. Additionally, having datasets that accurately represent the target population for various use cases is vital to reduce algorithmic bias.

This document has been formatted to be read conveniently on screens with landscape aspect ratios. Please print only if absolutely necessary.

Authors
The authors are researchers working with the High-Tech Geopolitics Programme at the Takshashila Institution.

AI development relies on computing infrastructure, especially in the cloud. Promoting competition, reducing entry barriers, and preventing monopolistic practices in the AI cloud service provider market is essential to encourage a competitive compute ecosystem. Developing domestic capacity in cloud infrastructure is also important for strategic autonomy and technological resilience in areas with national security implications.

At the model stage, a competitive marketplace that fosters innovation and accessibility is essential, offering various distribution models from proprietary to open-source. Governments should also support open-source AI technologies with broad research and commercial applications.

Lastly, addressing concerns in other stages of the AI ecosystem will pave the way for the market to address the requirements of the application layer effectively. However, it is essential to adopt a risk-based framework to help strike a balance between proceeding cautiously in high-stakes AI applications and encouraging innovation in other areas.

Acknowledgments

The authors would like to thank Pranay Kotasthane, Gangadhar Nittala, Kailash Nadh and Deepanker Koul for their valuable feedback and comments.

The authors also conducted a roundtable to discuss the ideas in this discussion document. They extend their gratitude to the following participants for their insightful comments which helped refine this work: Amlan Mohanty, Anand V, Deepak V S, Madhavan Mukund, Manjulika Vaz, Narayan Ramachandran, Prateek Waghre, Rahul Matthan, Rohit Satish, Saurabh Chandra, Shambhavi Naik, Vaneesha Jain.

Index of Abbreviations

AI	Artificial Intelligence
ASIC	Application-Specific Integrated Circuit
ASPI	Australian Strategic Policy Institute
CPU	Central Processing Unit
CSP	Cloud Service Provider
CUDA	Compute Unified Device Architecture
DEPA	Data Empowerment and Protection Architecture
DPDPA	Digital Personal Data Protection Act, 2023
EU	European Union
FPGA	Field Programmable Gate Array
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
GPT	General Purpose Transformer
GPU	Graphics Processing Unit
LLM	Large Language Model

ML	Machine Learning
NIST	National Institute of Standards and Technology
OECD	Organisation for Economic Co-Operation and Development
OTT	Over-The-Top services
PaLM	Pathways Language Model
RISC-V	Reduced Instruction Set Computer-V
RMF	Risk Management Framework
SoC	System on Chip
TPU	Tensor Processing Unit
UPI	Unified Payments Interface

Table of Contents

I. Introduction 8

II. The AI System Supply Chain 11

III. Values for AI Governance 14

Human Flourishing 14

Democracy 15

Stable Global Order 16

Competition 17

Planetary sustainability 18

IV. Desired Outcomes in the AI Ecosystem 19

Data 19

Ownership of Data 19

Unlocking Public Datasets 22

Computation 24

Model 26

Application 27

V. Insights Into the AI Ecosystem 29

Data 29

Proprietary Data Offers a Significant Advantage to Vertically Integrated Platforms 29
A Marketplace for Data 30
Copyright legislation 32
Computation 33
Potential Competition Concerns in Cloud Computing 34
Sovereign Computing Resources 35
Building Domestic Capacity 36
The Compute Conundrum 39
Sustainable Development 43
Models 44
Bridging the Talent Gap 44
The Real Risks of General Purpose AI Models 46
The Gradients of Openness in 'Open' AI Systems 47
Application 51
VI. Key Questions 53
VII. Appendix 55
A Framework for High-Technology Geopolitics 55
An Overview of The AI Chips Market 56
Characteristics of AI Chips Supply Chain 61
VIII. References 63

I. Introduction

The transformative potential of artificial intelligence has captured the world's imagination. It has been compared to other game changers of the past millennium, such as the Industrial Revolution and the invention of the Internet. Its importance cannot be overstated as a critical technology that can lead to hyper-growth in multiple downstream applications, including those with defence and national security applications.

AI systems have come a long way since their inception. Simple rule-based systems, where logic and explicit rules were paramount, transitioned into the machine learning era, where past data was used to make future predictions. The advent of deep learning powered by larger datasets and computing heralded a new era, leading to significant advancements in image[1] and speech[2] recognition.

Further advances, such as the development of transformer model architecture[3] coupled with improvements in computing and availability of massive amounts of data, have led to the more general-purpose artificial intelligence systems we see today.

General-purpose AI systems can be adapted to a wide range of applications, including those for which it was not intentionally and specifically designed[4].

The increase in capabilities is a natural progression in the capabilities of AI systems with the improvements in architectures and the exponential growth in data and computing. Their adaptability to various downstream applications makes them critical and has led to a global race to set the rules for AI[5].

The current discourse on the risks and potential of AI is largely driven by a handful of dominant technology companies and the media narratives they propagate. The global AI governance efforts are focused on minimising AI's harmful effects or preventing bias and discrimination in AI systems. While these concerns are important, they often overshadow immediate policy-related questions — the challenge of governing the rapidly evolving AI industry.

This document focuses on AI governance from a national interest perspective. It focuses on how AI adoption can help humans flourish, strengthen democracy, and promote a stable global order. It highlights the need for sustainable practices in AI adoption and the importance of ensuring competition across all stages of the AI ecosystem. The AI supply chain has multiple stages: data, computation, model, and application. Entry barriers exist at each of these stages. In some instances, many are vertically integrated and controlled by a single company. Competition is vital at the ecosystem level and is also desirable at each of these stages.

While acknowledging concerns around AI ethics and safety, which are the primary focus of governance efforts globally, this document adopts an

AI Hype and Doomerism

Key Silicon Valley figures, including OpenAI's Sam Altman and Elon Musk, have called for regulatory measures hyping speculative existential threats. Other experts, however, point out that the real risks are more immediate and relate to the concentration of market power and a lack of accountability.

industry governance approach that can help capitalise on the benefits of AI while mitigating the potential risks.

The recommendations in this document are equally applicable to purpose-specific machine learning systems and general-purpose AI systems. Although certain examples specifically mention general-purpose AI models or specific AI applications, the scope of the document covers AI systems in general.

II. The AI System Supply Chain

The development of AI systems has inputs such as data, computation, models, and applications as different stages or components in the supply chain. This applies to purpose-specific machine learning systems and general-purpose AI systems. These components can be visualised as layers, with data and computation contributing to the model, which, in turn, support the applications.

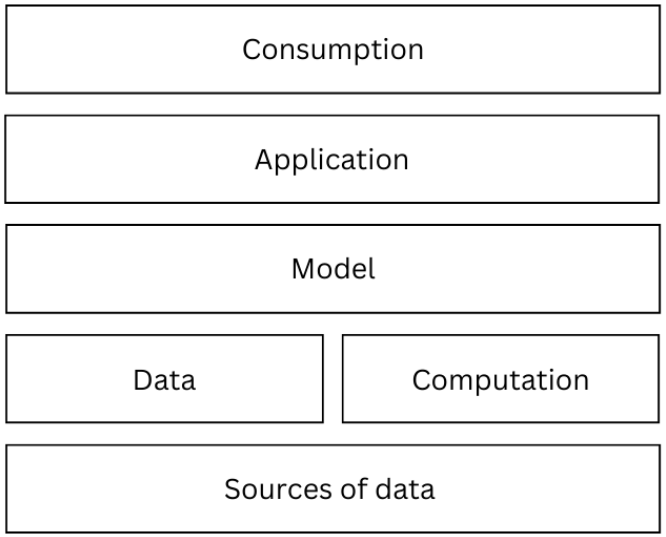

Figure 1: The components of the AI supply chain

While the ideal situation would be to have competition at each stage in the supply chain, in practice, many of these stages are vertically integrated and controlled by a single company. Vertical integration exists when a company controls more than one stage of production or distribution of a particular good or service. It could lead to the creation of entry barriers and discourage competition. Some scenarios are described in the examples below.

The supply chains for AI systems with different levels of vertical integration are illustrated below. These have been shown for general-purpose AI systems but are equally applicable for more specialised machine learning applications as well.

Scenario 1: Fully vertically integrated AI supply chain

This supply chain model features a fully vertically integrated entity that owns its computational resources, has access to both public and private proprietary data, and builds and deploys its own AI models and the applications that run on them.

Google exemplifies this level of integration. Their operations span across the supply chain, from having their own chips to access to vast amounts of data, including proprietary data. They also offer cloud services and have integrated their AI systems into various applications for both web and Android users.

A variation of this model is an entity that offers a fully integrated system as an "AI-as-a-Hub" service. Third-party models can be used alongside the entity's own models in this setup. Amazon's Bedrock serves as an example of this approach.

Scenario 2: Partially vertically integrated AI supply chain

This type of supply chain involves a tight coupling between some stages of the AI supply chain. AI developers typically rely on cloud computing resources for training and deploying their models. Most major AI companies have already formed strong partnerships with cloud providers. For example, Microsoft and OpenAI have chosen this approach. They train their models using Microsoft's cloud infrastructure and make them available exclusively through Microsoft Azure.

Scenario 3: Fully Disaggregated AI supply chain

Lastly, in a disaggregated model, we can envision a scenario where the computation, data, and AI models all originate from different players that are not interconnected through close partnerships. This supply chain model is often favoured by academia and other open-source AI developers who make their entire models available for replication and deployment. A recent example of this type of disaggregation in the cloud and computing space is Nvidia tying up with the likes of Oracle, Google, and Microsoft to offer its DSG supercomputer as a cloud service for building AI models.[6]

III. Values for AI Governance

For each stage of the AI supply chain, from data to application, the authors of this document have identified guiding values to ensure that AI not only advances technologically but does so in a manner that upholds societal and national interests, safeguards democratic values, and preserves our environment. These values are Human Flourishing, Democracy, Stable Global Order, Competition and Planetary Sustainability. We delve deeper into the significance and application of each value in the context of AI governance below.

Human Flourishing

Vijay Kelkar and Ajay Shah[7] propose that the toughest challenges for a state—such as the tax system—involve processes that feature a high number of transactions, the need for discretion, high stakes for individuals, and some degree of secrecy. AI adoption could reduce the complexity of such challenges on some of these dimensions, such as the transaction volume and discretion. This makes it easier to overcome state capacity limitations and deliver better governance and public services. AI is an effective tool to enhance human cognitive capacity and productivity. Its potential to disrupt multiple industries has been likened to the industrial revolution. However, there are also fears about the impact on jobs. An OECD report cautions that

27% of jobs are at a high risk of automation[8]. These could be in fields as varied as accounting, finance, or medical diagnosis. However, the path of technological advancement is not preordained. AI development can enable growth that is broadly distributed. In this context, there is a growing consensus for AI development to be focused on augmentation instead of automation. This approach aims to empower individuals and organisations by harnessing AI to enhance their capabilities rather than replacing human workers entirely.

AI and Bias
AI systems are now playing a growing role in determining hiring decisions, access to credit, and even law enforcement. However, due to the imperfect nature of such systems, they can lead to unfair outcomes, exacerbating historical inequalities or discriminating against already marginalized individuals.

Democracy

AI systems can quickly process and analyse vast amounts of data, increasing effectiveness and reducing opportunities for rent-seeking at various stages of the policy lifecycle, including design, implementation, monitoring, and evaluation.

How AI Impacts Democracy
AI-powered tools can directly damage public trust in democratic procedures. At its most benign, the automation of legal processes by AI can challenge the transparency and accountability essential to democratic legitimacy. At the other end of the spectrum, AI-driven social bots and deepfakes can sway public opinion by amplifying disinformation, and discrediting political opposition.

However, such powerful, opaque, and imperfect systems controlled by a few can lead to externalities that undermine democracy. Several concerns have been raised about the biases inherent in these systems and the potential for misuse[9].

AI systems could potentially be used to create an Orwellian state. The powers to surveil citizens, monitor communications, and stifle dissent at scale can be

disastrous for democracy. For instance, such systems have been deployed in China to extend the state's surveillance capabilities and maintain social control[10]. Malicious actors can also exploit AI systems to disseminate false information, eroding trust in institutions and potentially influencing the outcomes of elections.

The use and misuse of AI systems could seriously undermine justice, liberty, equality and fraternity. Ensuring such systems are built and used in ways that don't undermine democracy is paramount. Many minds across the world are busy identifying the risks that AI poses to democracy, but while doing that, we should also ensure that the benefits from AI are not lost to humanity.

Stable Global Order

This represents considerations for safeguarding a state's interests in a rapidly changing world order. As Pranay Kotasthane notes in his paper on high-tech geopolitics in the post-pandemic world[11], technology and geopolitics are increasingly getting intertwined. He observes that trade wars are likely to be tech wars at their core, private technology giants are expected to align more closely with their respective governments, and geopolitical factors will influence international tech collaboration. AI is a critical technology with many downstream applications. Ensuring a diversified supply chain to the building blocks of AI systems is critical to ensuring strategic autonomy.

AI systems are also increasingly being utilised in military and national security applications. Integrating AI into these areas could lead to an imbalance of power, favouring those who have access to these technologies over those who do not. In the realm of international relations, which often operates on the principle of amoral realism, it is crucial to prevent a situation where the power balance is excessively tilted in favour of a particular adversary.

Competition

The AI development supply chain has different stages, such as data, computing, models, and applications. Companies lacking access to any of these stages will struggle to compete effectively. Big tech companies can also leverage their existing market power and insights into user preferences and behaviour to gain a significant advantage in new markets. For large general-purpose AI systems, some stages will also have prohibitive costs, network effects, and economies of scale that benefit entrenched players.

Competition drives companies and individuals to innovate and improve. It increases accessibility and choice for end users and is more likely to cater to a broader range of societal needs. Therefore, an effective AI governance framework should create an ecosystem that enables competition at every stage of the supply chain. It should also guard against regulatory capture by early movers who have a big lead in developing AI models.

Planetary sustainability

The relentless drive among companies to expand the scale of AI systems has resulted in the development of computationally intensive models that depend on massive datasets. The accuracy of these models depends on exceptionally large computational resources that result in significantly high energy consumption[12]. In addition to the carbon emissions, they consume vast amounts of water for cooling, leading to water shortages[13]. The environmental impact of training and running such models should be considered in decisions regarding their governance.

IV. Desired Outcomes in the AI Ecosystem

Data

Ownership of Data

Large technology companies collect and commodify vast amounts of personal data in exchange for the free services they provide users. This model treats data as exhaust from consumption to be collected and used by firms. As a result, we end up with large silos of data controlled by dominant technology companies that employ this data as a moat that hinders competition.

Enabling markets that offer access to comprehensive data repositories on fair terms will be crucial in promoting competition in various aspects of the AI supply chain, such as application development and model development. The key idea behind this market is that data fundamentally originates from individuals, and they are the rightful owners of the data generated through their use of digital services and products. Any companies acquiring, exploiting, or selling this data should be required to obtain the owner's explicit permission.

Data is the fuel that powers the growth of AI models and applications. Indians consume nearly 20GB of mobile data a month, a three-fold increase from 2018 to 2022[14]. This is expected to more than double by 2024, with 5G adoption acting as a catalyst. Given this trend, it is clear that India is likely to be amongst the largest producers of data in the world. How India governs the production, consumption and monetisation of its data will impact the AI ecosystem around it.

On the principle that individuals should dictate the usage and distribution of their personal data, we can envision a data marketplace that has the following attributes:

- Individuals should be able to dictate if, how, and by whom their data can be used. While it is common practice for firms to aggregate and transform the data they capture for internal and external use, individuals typically lack the means to do the same. India's Data Empowerment and Protection Architecture (DEPA) and the Account Aggregator Framework built on top of it illustrate a consent-based intermediary system[15] to facilitate such a transaction. While it does not restrict firms from continuing with their existing data practices, it offers individuals some control over their data. For instance, it allows users to use their financial data from multiple sources to access various financial services for personal or business needs. A similar framework could also be applied to unlock the value of various other data locked

The Digital Personal Data Protection Act, 2023
The Digital Personal Data Protection Act 2023 in India regulates personal data governance, empowering individuals (Data Principals) with rights over their data and redefining business practices for responsible data handling by imposing new compliance requirements on Data Fiduciaries and Data Processors. Furthermore, the relaxation of data localization rules under the Act facilitates cross-border data flows, which influence how foreign players collect and utilise data as they deploy AI technology in India.

The Account Aggregator Framework
It facilitates the exchange of user data between financial institutions such as banks, insurance agencies, and mutual fund companies, based on user consent. This unlocks user data from silos and allows the market to find competitive solutions to address customer's needs.

in silos for specific purposes approved by the user.. While DEPA's design is not suited for training AI models, it is suitable for inference applications.

- Data portability to unlock value for users across markets and also prevent undesirable platform lock-in effects. For example, a user's health data from a fitness tracker can be shared with their chosen healthcare provider to facilitate the creation of a personalised health plan. Data request templates for different use cases need to be established for various use cases in consultation with stakeholders across industry and academia.

- Enable data provenance and usage transparency. A data aggregation framework also offers information about the origin and lineage of a dataset available within the market. In addition, it can also serve as an audit trail for the downstream models where a dataset is being used.

- Ensuring free and fair access to data is an important outcome to enable a competitive AI ecosystem. A consent-based data aggregation framework will allow firms to create consolidated profiles of users, unlocking data from multiple service providers such as telecom operators, OTT players and others. In addition, governments and large private players should be able to publish datasets to the marketplace. In cases where the data represents the flow of interactions between different users rather than something distinct about a single user, the

idea of collective ownership of data can also be implemented[16]. An example is New York City requiring ride-sharing companies such as Uber and Lyft to disclose data on the date, time, and location of pickups and drop-offs. The city intends to use all this data to understand traffic flow better and plan effectively.

This section proposes a rethinking of data ownership that will affect not only the use of data for AI training but also business models across various industries. Although it introduces considerable scope expansion, we present it as an alternative to the status quo.

Unlocking Public Datasets

The principles of "public money, public code" or "public money, public data" are applicable here. Taxpayer funds are used to create datasets with economic or research value, but these datasets are often kept inaccessible to the public. By unlocking this data, significant public value can be generated.

For instance, having data that accurately represents the target population helps reduce bias in algorithms. This is especially evident in fields such as medical diagnosis, where factors like race, gender, and lifestyle can impact disease likelihood[17]. Assisted diagnosis using image recognition algorithms needs large datasets of labelled images from patients with confirmed diagnoses. Creating such datasets will have positive externalities for research and the AI applications they enable. Additionally, archival datasets from

Doordarshan or All India Radio, rich in multimodal and multilingual data, could significantly enhance AI language tools across various languages. Governments should invest in making such representative datasets available for use in both research and commercial applications.

Additionally, government data often exists in isolated databases, lacking a structured data engineering plan and may not be in a suitable format for training AI models. Thus, there is a need for consistency in data collected from various sources. For instance, during the COVID-19 pandemic, data from healthcare, vaccination, and contact tracing needed to be uniform and available in real-time for designing effective response strategies[18].

Comments received by the authors from a roundtable seeking feedback on this document highlights that while the benefits of unlocking public datasets are recognised, there exist several gaps in proceeding with implementation. There is a need to connect data providers with consumers to arrive at common standards for sharing data – such as which data points are to be collected and at what granularity. There is significant effort involved in cleaning and curating datasets so they can be used for the necessary applications. Training the necessary workforce and establishing best practices for data sharing are also significant hurdles. Additionally, ensuring accountability for adherence to these practices is essential.

The authors also received feedback about the need for purpose driven collection of datasets to enable innovation. For instance, AI for health

applications relies on public datasets from the US or UK which might not be relevant for the Indian population.

The authors propose the creation of a sector agnostic entity staffed with data scientists and cybersecurity experts to ensure data uniformity and compliance with best practices. Its mandate would include creating a data engineering plan, conducting audits of different state entities, and enforcing data compliance standards. The mandate should be restricted to non-personal data, and anonymised personal data which will have minimal risks on privacy or data security for individuals.

Computation

Computing infrastructure, especially when delivered through the cloud, is a vital enabler for developing and adopting AI. It provides access to data storage, processing capabilities, and advanced analytics at scale. However, the use of cloud computing also presents challenges to the nascent AI industry. These challenges include potential anti-competitive behaviour by cloud service providers (CSPs), data security and privacy risks, and reliance on foreign-based supply chains for computing resources.

The Indian AI industry primarily depends on computing provided by foreign-based CSPs such as Microsoft Azure, Google, and Amazon Web Services[19]. There is a strong case for India to develop its domestic computing

capabilities to mitigate the risks of potential disruptions caused by natural disasters, human-made incidents, or global supply chain disruptions related to AI chips.

Dominant CSPs often have vertically integrated AI supply chains, meaning they might compete directly with their own cloud computing customers. This takes on greater significance in a market like India, where the existing industry mostly sits at the application layer. Even in the application layer, the industry heavily depends on AI models from industry giants like Google or Microsoft. This trend could eventually result in a concentration of market power in both the computing and application layers.

Therefore, the following outcomes need to be enabled:

- Foster a competitive and diverse market for AI cloud service providers by promoting fair competition, reducing entry barriers, and preventing monopolistic practices or vendor lock-ins.
- Promote investments in building domestic capabilities to participate in the global value chain for AI chips and essential computing hardware.
- Create sovereign computing resources for AI to cater to military and government applications while also serving the needs of industry and academia. This includes GPU/ASIC clusters, data centres, and networks.

Model

AI systems can drive growth in numerous downstream industries. A competitive marketplace for AI models is vital and can include a variety of distribution models, ranging from fully proprietary to fully open[20]. These models can operate on different business strategies, such as pay-for-access like OpenAI's ChatGPT, owning the innovation ecosystem like Meta's LLaMA, or providing a model that serves as a base for research and innovation, like BigScience's BLOOM.

For AI technologies with broad research and commercial applications, rather than reinventing the wheel, governments should recognise and support established open-source communities by providing grants[21]. These grants should be provided regularly to fund critical open-source AI projects, whether they are domestic or international in origin. Corporations have also embraced this funding approach, and it has proven to be a sustainable method for sustaining open-source projects. As Frank Nagle suggests for public funding of security support for widely used OSS projects[22], the authors believe that the tangible and intangible benefits in this case could also justify the investments.

Discussions from a roundtable on this document suggest funding open-source AI models for applications in areas where demand exists, but the population

cannot afford them. For example, using AI to screen for oral cancer targets a demographic that cannot afford such services, which the market would not typically address.AI systems designed for image recognition, autonomous driving, and text analysis can have dual-use applications and are important in defence and national security contexts. These systems can operate in hazardous environments and enhance data processing and decision-making capabilities, substantially improving military capabilities. For instance, China has been actively pursuing the development and integration of AI in various military applications as a part of its strategic efforts[23].

Restricted access to such versatile AI technologies, crucial for both economic prosperity and military purposes, poses a significant vulnerability. Thus, in addition to having a vibrant AI marketplace, to the extent that such technologies are dual-use, it is essential to ensure strategic autonomy in access to such systems. Increasing investments in private and public research and development (R&D) is necessary to achieve these objectives. Additionally, establishing centres of excellence that can attract and nurture top AI talent will be essential in reinforcing domestic AI capabilities.

Application

As machine learning and AI models continue to improve, applications will increasingly rely on them to deliver new and innovative features. While AI can enhance cognitive capacity and productivity across various sectors, there

are also risks in the use and misuse of AI that could seriously undermine democracy.

AI holds the potential to substantially increase productivity across diverse industries such as education, transportation, agriculture, finance, and customer service. It can also improve transparency and efficiency within government, improve outcomes for citizens and reduce opportunities for rent-seeking.

However, there is a darker side to AI as well. AI systems could potentially be exploited to create a surveillance state, enabling governments to monitor citizens, surveil communications, and suppress dissent more effectively. Additionally, AI systems can be used to spread disinformation and micro-target political advertisements, undermining the functioning of democracy.

Thus, it is essential to adopt a risk-based framework that takes into account the nature of the AI system and the sectors in which they are proposed for use. Adoption of a framework similar to the NIST AI Risk Management Framework[24] can better manage the risks associated with AI systems in civilian or government applications. This approach can help strike a balance between proceeding cautiously with high-stakes AI applications while also encouraging innovation.

V. Insights Into the AI Ecosystem

Data

Proprietary Data Offers a Significant Advantage to Vertically Integrated Platforms

Vertically integrated technology firms often have easier access to proprietary data, giving them a significant edge in developing AI systems. Such proprietary data can include social media interactions, code, academic publications, books, and insights into user behaviour.

For instance, a video conferencing platform, such as Zoom, can harness customer content generated by its vast userbase to introduce AI-enhanced features, like meeting summaries, thereby enhancing its product[25]. Similarly, web-based software development platforms like GitHub can tap into their vast code repositories to provide tools such as GitHub Copilot, substantially improving user productivity[26]. Search engine providers have an advantage in accessing publicly available data on the Internet as they would have built a web index, a sorted and categorised index of web crawl data, which allows them to provide accurate search results rapidly[27]. Web crawlers of search engines are also less likely to be rate-limited as websites would want to appear in search results, giving companies like Google and Microsoft an advantage in accessing publicly available data[28].

Models trained on high-quality data tend to perform better[29]. This includes data from sources such as books, academic papers, news articles, and Wikipedia. Studies estimate that for training large language models, such data is likely to be exhausted by 2027. Thus, access to proprietary data can be a key differentiator. Proprietary data may be purchased from different sources, but vertically integrated platforms will have a significant advantage due to easier access to proprietary data.

Big tech platforms tend to be monopolies or duopolies due to the network effects, which increase the utility of the platforms for users. The scale and insight into user behaviour help these companies innovate better than competitors. This market consolidation is evident across various platforms, ranging from search engines and social media to ride-sharing and food delivery services. Proprietary data is valuable not just to improve the performance of the AI systems but also to improve their offerings by integrating these AI systems. This helps them further consolidate their market power and will be especially useful for vertically integrated platforms.

A Marketplace for Data

We have become accustomed to the idea of a free online experience. Services such as email, messaging, calling, social media, or video sharing are all enjoyed without direct monetary costs. Users expect these services for free and are not paid for the data they generate on these platforms. The race to capture and monetise the time and attention of users has led to the widespread collection

and commodification of personal data by corporations, what Shoshana Zuboff has termed surveillance capitalism[30].

Jaron Lanier argues in his book that although the exchange seems like a barter — free data for free services — it's problematic. He argues that this approach distorts traditional market evaluation principles, unfairly distributes financial gains from the digital economy, and prevents users from developing themselves into "first-class digital citizens"[31].

With data being locked in silos controlled by big tech companies, users cannot monetise their data, and other companies cannot compete effectively against these giants. Arrieta-Ibarra et al.[32] compare this to an extreme version of a monopsony (a market where there is only one buyer and who, therefore, has control over the negotiations), where users are not even aware of the value of their data.

Regulatory frameworks, such as the European General Data Protection Regulations and India's Digital Personal Data Protection Act, increasingly recognise the ownership rights of the users who generate the data. Creating a market for data and unlocking of user data from these silos can generate huge social and economic value.

India's Data Empowerment and Protection Architecture (DEPA)[33] is founded on the principle that individuals should dictate the usage and distribution of their personal data. This aims to offer Indians the chance to

enhance their own well-being through control over their data. The Account Aggregator Framework is an implementation of this architecture for the financial sector and has the potential to unlock massive value for businesses and end users.

Copyright legislation

In the United States, the fair use of a copyrighted work is allowed based on the following four factors[34]:

- The purpose and character of the use, including whether such use is commercial or is for non-profit educational purposes.
- The nature of the copyrighted work (facts cannot be copyrighted, but creative works can).
- The amount and substantiality of the portion used in relation to the copyrighted work as a whole.
- The effect of the use upon the potential market for or value of the copyrighted work.

For instance, Google Books scans entire books for their search engine, but a select few pages are displayed to a user. Courts have ruled that this is not a violation of copyright as the purpose and character of the use are different[35].

In other words, searching through a book is a transformative use case as opposed to reading the book in its entirety.

Multiple ongoing legal battles allege that generative AI systems have infringed on copyrighted works. For example, Microsoft and OpenAI are facing a class action lawsuit alleging that the AI-powered coding assistant, GitHub Copilot, indulges in "software piracy on an unprecedented scale"[36]. Stock photography firm Getty Images is suing Stability AI, the creators of the open-source text-to-image tool Stable Diffusion[37]. Authors Mona Awad and Paul Tremblay have also filed a similar lawsuit against OpenAI[38].

The law on these matters is still evolving, and a decision favouring the plaintiffs could reduce the data available for training or lead to higher costs to license such data. These outcomes would favour vertically integrated companies with access to large proprietary databases. This could also pave the way for powerful intermediaries that bundle data licences without adequately compensating the original data creators[39].

Vertical Integration in the Cloud Services Space
Cloud Service Providers like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform exhibit vertical integration by controlling multiple stages of their service offerings. For example, AWS not only provides cloud infrastructure but also develops its own hardware, such as the Graviton processors, and offers a wide array of services ranging from storage and computing power to machine learning and analytics tools. This integration allows them to optimize performance, reduce costs, and quickly innovate, but also raises concerns about market dominance and competition.

Computation

Currently, there are two main ways to obtain computing infrastructure. One option is for end-users to buy and set up their own hardware, which can be quite costly. The other option is to use cloud computing services, where

providers offer a package of services. The latter is especially alluring for developers or end-users, as CSPs are willing to provide computing resources at deep discounts for AI research and development to firmly entrench their market share as the industry grows.[40]

Cloud computing is more than just a technological infrastructure; it is a powerful tool that can break down barriers to accessing technology and essential digital services[41]. This levels the playing field and reduces barriers for innovation. Increased competition in the AI space, enabled by cloud services, leads to more options and improved services for consumers.

Potential Competition Concerns in Cloud Computing

As mentioned earlier, it's crucial to maintain competition among different cloud computing service providers to ensure affordable and accessible computing resources for the nascent AI industry. The primary sources of these services for a country like India are foreign-based CSPs and hyperscalers. Domestic hyperscalers have not yet reached a similar scale in their operations. One of the main obstacles they face includes dealing with infrastructure challenges that are common in developing countries, such as ensuring a consistent and reliable power supply[42].

The leading CSPs, such as Google, Amazon, and Microsoft, not only offer computing services but also compete in different stages of the AI supply

Cloud Egress Fees and Customer Switching

Egress fees charged by cloud service providers for data transfer out of their networks can significantly impede customers' ability to switch providers. These 'hidden' fees act as a deterrent against leaving a cloud provider's ecosystem, as they can substantially increase over time, adding to the overall cost of cloud management. While data ingress (moving data into the cloud) is typically free, egress charges apply to data moving out, which can accumulate rapidly, thus elevating IT costs and creating a financial barrier to switching providers. These fees can escalate the costs of managing AI applications and datasets, especially because large volumes of data are involved.

chain, including models and applications. These CSPs, given their dominant position in the cloud computing market, may hinder competition and innovation in the AI ecosystem. They could do this by imposing restrictive contracts, imposing high switching costs, or unfair pricing on their customers or competitors. For instance, they might charge high egress fees for transferring data out of their platforms, making it challenging for customers to switch to different cloud providers[43].

Additionally, CSPs could use their access to vast amounts of data and advanced analytics capabilities to gain an unfair advantage in the AI market. They might achieve this by creating or improving their AI products and services or by acquiring or partnering with other players in the AI supply chain. For example, they already leverage their cloud platforms to gather and process data from various sources like e-commerce, social media, or IoT devices and use this data to train and improve their own AI models[44].

Sovereign Computing Resources

CSPs can also pose a risk to a nation's data security and sovereignty when they store and process sensitive data of citizens or entities on servers located in other countries, which may be subject to different legal and regulatory rules. For instance, CSPs could be compelled to share or disclose such data with foreign governments or agencies, and they might also be susceptible to cyberattacks or sabotage by malicious actors[45].

Keeping Pace with Generational Uplift in Compute
Upgrading GPUs is essential for AI developers and cloud service providers, as exemplified by the progression from NVIDIA's Turing to Ampere architectures. The A100 GPU, built on the Ampere architecture, offers substantial improvements in terms of AI performance and efficiency over its predecessors. These improvements include enhancements in floating-point operations, memory bandwidth, and support for newer technologies like TF32 and mixed precision training, which are highly beneficial for AI and deep learning tasks. This leap in capabilities is crucial for handling more complex AI models and larger datasets, demonstrating why staying current with GPU technology is critical for competitive AI development and cloud service provision.

An interim measure seeking to bolster a country's nascent AI industry to build and train its own models can be to create a publicly accessible supercomputer[46]. This involves acquiring a large number of GPUs and allowing start-ups in AI and other emerging technologies that require substantial computational power to rent time slices on this supercomputer[47]. However, there are some uncertainties associated with this approach. Firstly, current-generation GPUs like the Nvidia H100 are anticipated to be in short supply until mid-2024 due to high demand from almost every player in the AI space. Assuming it is possible to purchase enough GPUs, a generational uplift from the next generation of GPUs will enable faster AI and ML workload processing for those who have them. Time to market is an important consideration for AI companies, and this may be compromised if the only reliable access to computing happens to be slower than what their competitors may have access to.

Building Domestic Capacity

In light of these considerations, it is important to regulate CSPs to achieve the twin objectives of enabling unfettered access to cutting-edge computing resources while building strategic autonomy in computing.

At this point in time, where we are still exploring the capabilities of this technology and its ramifications, regulation pertaining to CSPs should adopt a light-touch and enabling approach. This approach should promote competition, diversity, and interoperability while ensuring compliance with

data protection and security standards. This could be done by creating an industry-led body responsible for setting and enforcing codes of conduct, best practices, and standards for cloud services, as recommended by the Telecom Regulatory Authority of India[48].

As cloud computing becomes increasingly integrated into various aspects of life, a country like India can leverage its sizeable market and negotiate favourable terms and conditions with foreign-based CSPs. This negotiation can encompass aspects like data localisation, taxation, dispute resolution, and liability clauses. This could be done by creating a common framework that can address the legal and regulatory issues arising from cross-border data flows and jurisdictional conflicts. This framework should also incorporate principles of data sovereignty, data minimisation, data portability, and data security, harmonising provisions from legislation like the Digital Personal Data Protection Act of 2023[49].

Another way to ensure access to compute resources and build strategic autonomy is to invest in domestic production and development of computing resources for AI, such as chips, servers, data centres, and networks. A governance framework should also support research and innovation along with front-footed policy measures such as sandboxes in emerging technologies such as quantum computing, neuromorphic computing, and edge computing, which are already used in current AI systems or may find use in future AI systems. These measures must be complemented by industrial

and trade policy efforts to encourage the involvement of domestic CSPs in the cloud market. This can be achieved by facilitating partnerships with global players[50] and offering incentives like tax breaks, subsidies, and preferential procurement policies. As of writing this, most jurisdictions worldwide have taken a wait-and-watch approach to address these potential governance challenges[51].

NVIDIA's early mover advantage and investment in CUDA has continued to pay off dividends as their hardware and software libraries are the de facto options in scientific and parallel computing applications, and consequently led to their near-monopoly in GPUs[52]. A sovereign compute infrastructure strategy that aims to procure chips will continue to remain dependent on NVIDIA because the alternative of building smaller, more efficient and application-specific chips will remain on the sidelines due to their lack of access to the CUDA platform and its libraries. Intel and AMD, which are the other major players in the GPU chips space, can have their chips run applications built using CUDA by running translation layers (ex: ZLUDA) but the use of such translation layers continues to be prohibited in NVIDIA's End User License Agreements.[53]

This is just one way in which NVIDIA seeks to maintain its technological moat, and exercise monopoly power. Amidst constrained resources, policy should ideally focus on seeking a course of action that creates a degree of invulnerability to market concentration like this. In India's case, perhaps this

may take the form of initiatives that leverage its comparative advantage in having a large workforce of globally proven chip design engineers, and invest in creating software emulation layers on top of which algorithms for parallel computing, machine learning and other AI workloads can be developed. However, further study is needed to formulate more concrete avenues of achieving this invulnerability.

The Compute Conundrum

Apart from these, the size of models that a state might want to build as part of its national strategy is a key consideration that can dictate the strategy to build sovereign computing resources and the extent of its domestic capacity, Industry voices seem to be leaning towards model development focused on solving specific use-cases (agriculture, healthcare, education etc.,) and therefore perhaps smaller custom models trained on relevant datasets are the path forward.[54] This could be a viable way in which emerging economies may not need to compete in the global arms race for either model or top of the line compute resources.

On the other hand, if there are indications that there are clear geopolitical advantages to building an AGI, then perhaps the imperative for building larger models for a state will emerge, and consequently, the need to build out larger and more advanced compute capacity.

The "**compute conundrum**" that states, therefore, face, is whether they should maximise their compute capacity to achieve the goals of AGI or just optimise their existing compute capacity for specific use-cases. A prerequisite to answering this conundrum is to have a reliable measure of existing compute capacity. However, in the case of India, and other emerging economies, no such metric is available. Combined with the lack of an accurate estimate of projected compute demand in the immediate future, it is a difficult endeavour to form an informed policy decision. As they attempt to resolve this quandary, policymakers should consider the following. **First**, domestic compute infrastructure must be scalable and flexible in order to accommodate shifting national digital priorities over time. **Second**, security and sovereignty considerations must be tackled keeping in mind that collaboration with friendly partner countries is essential for participating in the global AI market ecosystem. **Third**, building large models is an extremely energy-intensive endeavour, and therefore, understanding the implications of building compute capacity on a country's broader environmental goals is critical before formulating compute strategy.[55]

As of now, in order to attempt building AI systems focused on solving specific social or governance problems in a country like India, smaller, more bespoke models are needed, and therefore, a more targeted and modest domestic compute capacity.

Interestingly, in what has been hailed as an milestone for cloud compute in India, People+AI (an initiative of the non-profit EkStep Foundation) launched its **Open Cloud Compute (OCC)** project in May 2024.[56]

The OCC, in simple terms, is a digital public infrastructure approach to compute that aims to create an open and decentralised micro-cloud computing grid/network. This network would bring together independent providers (of apparently disparate compute resources) on a single availability marketplace, improving discoverability and access to computing power and related services from any provider. The project seeks to democratise access to cloud computing infrastructure, making it more resilient, interoperable, and adaptable for innovation. It would be useful to think of it as an equivalent of the Open Network for Digital Commerce (ONDC),[57] for micro data centres across India, whose services can be chosen by a client on the basis of required processing power, latency needs, geographical proximity, etc. So far, a 24-member consortium, including giants like AMD, Oracle, and Dell, alongside startups like Von Neumann AI and Vigyan labs, have signed up as partners on this project.[58]

While the technical, administrative, and implementation details of the OCC are currently unclear, there are a few ways in which this project can effectuate or affect the desired outcomes we envision for the compute stage.

First, the OCC and its concomitant DPI approach align broadly with the objective of promoting fair competition and reducing entry barriers. The

presence of a shared, decentralised platform means that the dominance of a few large cloud service providers can be checked by a "*network of smaller, interoperable micro-players that collectively behave like a mega player*." Alongside the reduction of entry barriers, this may also mitigate the risk of vendor lock-in.

Second, it is arguably a more cost-effective way to develop India's domestic computing capabilities by integrating local providers into the global compute network. While this cannot completely supplant the need for some sovereign computing resources essential for military and governance needs, etc, it has the potential to reduce a near-complete dependency on foreign-based cloud service providers

Third, the OCC's emphasis on "open" standards and interoperability may strengthen data localisation and data protection norms, reducing the risk of foreign interference or cyberattacks. Decentralised networks are generally more resilient to security and data breaches, further strengthening a push towards strategic autonomy.

However, in a bid to counter market concentration risks in the compute segment of the AI value chain, OCC can run into the risk of being run the same way the NPCI manages the UPI. The NPCI has a monopoly over the UPI architecture and often interferes in the UPI ecosystem to "*prevent the dominance and mitigate the systemic risk of failure of any single player*" by dictating limits on the number of transactions for third-party apps.[59] Having

a monopoly on a decentralised network means that efficient resource allocation and management can often lead to misguided policies like the above, which can distort market incentives. Even so, if the OCC becomes dominated by a few large cloud players (where economies of scale and existing customer bases matter a lot), market concentration risks may crop up again, similar to how large providers like Google Pay and PhonePe dominate the UPI platform.

While still in its early stages, the OCC is apparently designed to be used globally and will not just be limited to Indian customers. Questions abound regarding fiscal sustainability, who maintains the infrastructure, whether AWS and Google, etc., can also plug into this grid and offer their existing solutions, pricing, etc. The OCC's development merits further study.

Sustainable Development

As mentioned earlier, the current environmental cost of cloud computing services is significant [60]. These services require extensive physical infrastructure, including servers, data centres, and cooling systems, all of which have substantial environmental footprints. The energy consumption of these data centres is immense, primarily driven by the need to power and cool the vast arrays of servers. Attempts to govern an AI industry must attempt to strike a balance between the benefits of cloud computing and planetary sustainability. For instance, states can provide incentives for CSPs to plan server farms that rely on renewable energy like solar and wind

power[61]. Alongside addressing the sources of energy, infrastructure and server efficiency[62], it should inform future efforts to decarbonise cloud computing.[63]

Models

Bridging the Talent Gap

Developing cutting-edge AI systems demands substantial resources in terms of data, computing power, and technical expertise. The industry can mobilise these resources better than academia, and the recent breakthroughs in AI research indicate this - Google's paper on transformer models[64] and Microsoft's paper on Low-Rank Adoption[65]. This trend is underscored by the Stanford AI Index report, which reveals that in 2022, industry entities produced 32 significant machine learning models, in stark contrast to academia, which contributed only three[66].

Data and computing are vital components in the training of AI systems. We delve deeper into the challenges and obstacles related to these domains in the sections focused on data and computing.

In their paper discussing India's AI potential, Chahal et al. highlight some key observations[67]. India produces nearly twice as many master's level

engineering graduates as the United States, second only to China in this regard. However, India significantly lags behind the United States in producing PhDs, with less than one-third of the number. The shortcomings within India's higher education sector limit its ability to offer extensive training for a highly skilled AI workforce. Consequently, many Indian students opt to pursue PhD programs in foreign countries. The ASPI critical technology tracker clearly shows the brain drain of Indian AI talent to other countries, notably the United States[68].

Indian researchers publish AI-related papers at a prolific rate, trailing only behind the United States and China. However, when ranked by H-index, which measures both the productivity and citation impact of publications, India descends to the 16th position, indicating that the quality of these publications falls short of expectations[69].

India's research and development (R&D) investment is a mere 0.64% of its GDP, significantly lower than that of other nations[70]. Of this amount, approximately 37% comes from the private sector. In contrast, China allocates 2.4% of its GDP to R&D, and most developed countries devote more than 2% of their GDP to research and development. Switzerland, which topped the Global Innovation Index in 2022, spends 3.19% of its GDP on R&D. This implies that India's scientific research in AI is likely underfunded compared to many other countries.

On a positive note, according to a report from Bain & Company, India stands out as a significant global source of talent in data and AI skills[71]. It contributes 16% of the world's AI talent pool, ranking it among the top three talent markets worldwide. However, not all AI talent is equal. Top-tier AI researchers are involved in creating intellectual property or designing and training AI algorithms, which are activities at the top of the value chain. A study by MacroPolo, a US-based think tank, finds that over 80% of India's top-tier AI researchers move out of the country[72]. Consequently, while the specialised skills required for research and training AI models may be in shorter supply, India still holds a substantial advantage in engineering, which is likely also an advantage in developing applications based on AI models.

Comments from a roundtable on this document highlight the importance of active collaborations between industry and academia, similar to those in U.S. universities. With industry providing the data, infrastructure, and funding, and academia providing the talent, such partnerships can be mutually beneficial and highly productive. The participants of the roundtable also opined that the out migration of talent is not restricted to AI, but is also prominent in some other sectors. The Indian private sector's low investment in R&D also limits the opportunities available to talented individuals.

The Real Risks of General Purpose AI Models

A handful of companies, such as Google, Microsoft, OpenAI, and Meta, have a big lead in developing general-purpose AI models. A race is on to build

these powerful AI systems and lock in the early mover advantages. Top executives from Microsoft and OpenAI have called for an agency to regulate AI and licensing requirements to operate the most powerful AI tools[73].

An open letter by the Future of Life Institute supported by over 1,000 researchers, technologists and public figures has asked for a 6-month pause on training language models "more powerful than" GPT-4[74]. The letter presents risks of malicious use, job impact, and existential risks as serious consequences of embracing AI. However, as the authors of the book project AI Snake Oil point out[75], the letter exaggerates hypothetical risks and ignores the real issues around over-reliance on inaccurate tools, centralisation of power by these companies, and near-term security risks. An effective governance framework for AI should be able to address the significant challenges that come with AI adoption and not fall prey to the narratives peddled by various interest groups.

The Gradients of Openness in 'Open' AI Systems

The release practices of 'open' AI systems differ significantly from those of open-source software. There are varying degrees of openness in how AI systems are released. A study conducted by Radboud University researchers reveals the large variation in the availability, documentation, and accessibility across different AI models[76]. Additionally, unlike open-source software, significant access barriers in data and computing exist even in the fully open models.

Irene Solaiman's paper[77] introduces a framework for grading the openness of generative AI systems. Generative AI systems are a sub-type of general-purpose AI models that generate content based on user inputs, often across different modalities such as text, images or video. The framework classifies them into six gradients of access: fully closed, gradual or staged access, hosted access, cloud-based or API access, downloadable access, and fully open. This classification helps to understand the extent to which these systems are accessible to users and developers.

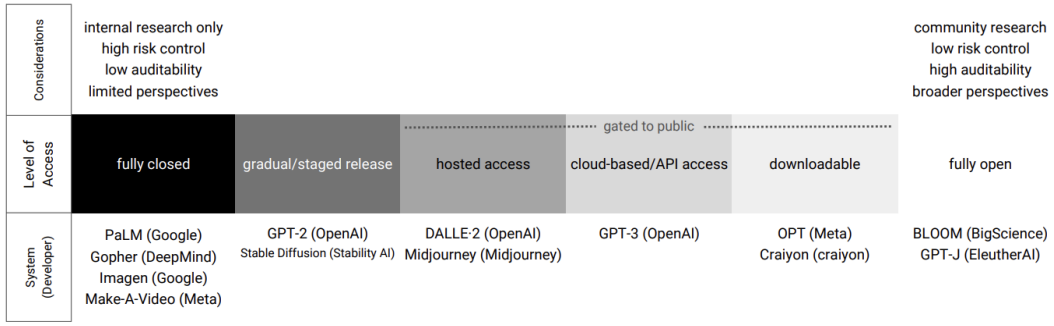

Figure 2: Source: The Gradient of Generative AI Release, Solaiman, 2023

With the fully open models, controls against misuse will be harder to enforce. However, they provide the reproducibility and independence from corporate decisions that are necessary for research purposes. One of the most widely used fully open models is BLOOM[78]. It is a multilingual language model

built by over 1,000 researchers from 70+ countries to overcome the access barriers that academia, nonprofits or research labs face to create, study, and use LLMs.

However, for most downstream applications, the various levels of convenience, customisation, ownership, and safeguards against misuse offered by hosted access, API access, or downloadable models will prove to be satisfactory. The model type selection will ultimately hinge on the trade-offs among these diverse criteria.

Another important consideration is that big tech companies also have vested interests in 'open' AI development. As a leaked memo by a Google researcher points out[79], "owning the ecosystem" is extremely valuable. This strategy is similar to what Google has done with Chrome and Android. He states that "by owning the platform where innovation happens, Google cements itself as a thought leader and direction-setter, earning the ability to shape the narrative on ideas that are larger than itself".

This is true of the dominant AI development frameworks PyTorch and TensorFlow, developed by Meta and Google, respectively, both of which are open-source. These companies continue to maintain them, and most AI models are trained on one of these frameworks[80]. Meta's downloadable model, LLaMA,[81] is also an effort at "owning the ecosystem".

Is Bigger Always Better for AI Systems?

Large language models have seen a massive increase in size and training data in the pursuit of better performance. Leading models such as PaLM and GPT4 use hundreds of billions of parameters and are trained on vast and varied datasets[82]. Bender et al. have raised concerns about the dangers of these massive models, coining the term "stochastic parrots"[83]. They point out the huge environmental, financial, and opportunity costs of pursuing research in a technology with many risks involved.

However, questions remain about the long-term sustainability of continually increasing model sizes:

- Limited availability of extensive high-quality datasets, diminishing returns from scaling up model size, and constraints on computing resources might affect the motivation or capability to develop even larger AI models[84].
- The costs of operating bigger models might be too high for most users, leading to a preference for more compact models. Market trends already indicate a focus on reducing the total cost of ownership of these models[85].

- In many business scenarios, an acceptable performance might be sufficient without needing cutting-edge accuracy.

Purpose-specific machine learning models might be more reliable and equally effective for many applications. However, if the performance of general-purpose AI models scales with size and data, it could lead to the wider adoption of larger models. As a result, building these models will likely become the domain of a few large players.

Application

Addressing concerns in other stages of the AI ecosystem will pave the way for the market to address the requirements of the application layer effectively. Competition issues at the data, computation or model stages could hinder innovation at the application stage.

For instance, ensuring unfettered access to diverse datasets can empower smaller players to compete effectively with dominant firms that have access to proprietary datasets. Regulating cloud service providers can prevent dominant firms from abusing their market power to favour their own applications over others. By making compute resources affordable and accessible, smaller firms can innovate on par with larger players. Finally, a competitive market for AI models will lead to reduced costs, increased innovation, and more choices for application developers. When concerns in

other stages are properly addressed, it will enable a wider variety of applications catering to diverse needs.

In addressing the AI application layer, we emphasise a risk-based governance approach. This involves identifying potential risks such as data privacy, security, and ethical concerns and incorporating trustworthiness into AI design and use. Compliance with global standards, like the NIST AI Risk Management Framework[86], can guide this process.

Regular stakeholder engagement is crucial, including feedback from developers, end-users, and regulators. This ensures the framework remains relevant and effective. Additionally, periodic reviews and updates will align the governance model with evolving technologies and market dynamics, fostering a competitive and innovative AI application landscape.

VI. Key Questions

- Big tech companies that develop AI models operate on a global scale. Is there a valid argument for pursuing multilateral harmonisation of AI governance efforts to achieve the desired outcomes?

- Given the notable market concentration in computing and the prohibitive costs, is there a valid argument for developing a domestic AI cloud infrastructure for use by both industry and academia?

- Unlocking data from silos can spur innovation across various industries. For instance, enabling access to fitness and healthcare diagnostic data can facilitate advancements in personalised medicine. Can a framework like the Data Empowerment and Protection Architecture (DEPA) be utilised to unlock the value of this data for consumers?"

- Government datasets are being made accessible through portals such as https://data.gov.in/. What improvements can be implemented to enhance access to research and innovation using this data?

- What considerations should be taken into account when deploying AI systems for e-governance applications?

- What initiatives are required to develop AI systems capable of addressing use cases specific to India, considering its high diversity, significant demand for low-skilled jobs, and limited state capacity?
- India loses its best talent to other countries. What efforts are necessary to reverse this brain drain?

VII. Appendix

A Framework for High-Technology Geopolitics

Governments worldwide are now deeply invested in securing their access to critical and emerging technologies. It's not just the domain of technology or finance ministries; national security experts and geopolitical policy analysts also focus on it. In a working paper, Pranay Kotasthane highlights several key trends[87]:

- Trade wars are likely to be tech wars at their core
- Aggressive national competition over high technology might produce some nonlinear breakthroughs this decade
- There will be higher alignment between private high-technology players and their national governments
- We will likely encounter selective international cooperation on high-technology subject to geopolitical considerations

The paper also introduces a framework to understand how nations might use political and economic tools to achieve strategic goals in high-tech sectors.

Assumed Impact on National Power	Strategic Objective	Instruments Used	Underrated Repercussions
Technology X underlies another critical & emerging tech	Denial	Secondary Sanctions	Difficult to sustain; incentives for backroom deals with adversary
		Restrictions on the movement of high-tech labour	Can slow down technical progress
		Export controls, End-use restrictions	Encourages adversary to build local capacity in a focused manner
		Investment restrictions	Can slow down technical progress
	Outpace adversary	Industrial espionage to steal secrets, targeted poaching	Invites stricter controls on professionals from the stealing country
		Build partnerships for resilience	Self-sufficiency is a myth.
		Indigenisation and industrial policy	Difficult to sustain.
		Sabotage	Self-damage
		Increase dependence and control	Helps manage the adversary's pace to an extent
	Remove major bottlenecks	Promote Open Source	Still a nascent field
		Build partnerships	Self-sufficiency is a myth.
Technology X can have direct cognitive effects	Influence minds and actions	Espionage	Limited impact on national power
		Decouple information flows	
		Disinformation	

Figure 3: A framework for High-technology Geopolitics. Source: The Takshashila Institution[88].

An Overview of The AI Chips Market

Development and deployment of general-purpose AI models and other AI/ML applications require high-speed and parallelised calculations that conventional general-purpose CPUs cannot perform well. At the same time, because of the large amount of data required to train the models, extremely

fast and high bandwidth memory also needs to be part of the computing process. Specialised memory utilised in AI accelerator chips/GPUs offers more than 4.5 times the bandwidth of conventional memory.[89] AI developers and machine learning researchers take advantage of Graphics Processing Units (GPUs), which offer both these capabilities (originally intended for image processing) to speed up computational tasks.[90]

Alongside GPUs, a set of AI accelerator chips is also designed for specific kinds of AI/ML and deep learning workloads. Both types of accelerator chips are deployed in hundreds of numbers for pre-training foundational models.[91]

Compute resources are required at two stages of the layer-based model we propose: at the training/development level and, subsequently, the inference/application level.

The number of parameters of a particular model determines the extent of the compute resource required for pre-training and inference.[92] Compute costs and power requirements increase exponentially as the size of models grows.[93]

Model	Parameters	Training data (in tokens)	Training time (in days)	Hardware (GPUs/TPUs)
LLaMA[17] (Meta)	65B	1400B	21	2048 A100 GPU
LaMDA[18] (Google)	137B	2810B	57.5	1024 TPU v3
GPT-3[19] (OpenAI)	175B	300B	34 [estimated]	1024 A100 GPU [estimated]
MT-NLG (Microsoft/ NVIDIA)[20]	530B	270B	90	4480 A100 GPU

Figure 4 Training data, training time and hardware used for training for a selection of models of various sizes.

Figure 4: Source: UK CMA AI Foundation Models: Initial Report[94]

The compute costs for training models are not public, but it is estimated that the largest Foundation model on the market (GPT-4) costs USD 100 million to train.[95] Inference costs are estimated to be around USD 700,000 a day.[96] Developers who are not already vertically integrated and want to avoid bearing the huge upfront costs required to build the computation infrastructure prefer to contract the services of CSPs. CSPs allow access to both general-purpose and AI/ML workload-specific computing resources, including CPU, GPU and storage, on contractual terms through the cloud.

The core of the market share for AI accelerator chips belongs to Nvidia, which accounts for 91.4% of the enterprise GPU market.[97] Both vertically

integrated developers of AI models and CSPs purchase Nvidia GPUs for AI-specific workloads. The reason for this huge concentration of market power is the proliferation and development of CUDA,[98] a proprietary general-purpose computing platform and programming model owned by Nvidia that allowed developers to make GPU-specific applications.[99] Since its inception in 2007, its ease of use saw its adoption in teaching curriculums across universities worldwide, as well as extensive use in the scientific research sector, which leveraged parallel computing offered by GPUs, which were relatively cheap when compared to renting supercomputer services. The network effects of CUDA becoming the de facto standard in programming to program GPU-specific applications meant that researchers tended to be locked into using Nvidia's chips as well since CUDA is difficult to port.[100] The closest competitor, AMD, has lower-priced offerings but, so far, has not been able to effectively combat Nvidia's first-mover advantage and technological lead.[101]

There is also a qualitative difference between compute requirements needed for training and inference-based workloads, respectively. The former's considerations of accuracy, ability to crunch large datasets parallelly, and training speed require significantly more raw computational power, memory capacity, and networking capabilities deployed over a large number of nodes (number of chips), which are utilised to nearly 100 per cent for long periods of time.[102] This also necessitates significant cooling capacity and high sustained power draw. Training workloads generally leverage GPUs,

manufactured using cutting-edge fabrication technology,[103] with Nvidia being the market leader in this segment.[104] This segment reportedly accounts for around 20% of the demand for AI chips.[105]

Compute infrastructure needed for inference-based workloads prioritises latency and throughput as opposed to raw computational power and hence needs high-capacity fast memory and high bandwidth I/O channels.[106] Most inference workloads are performed on traditional CPUs, supplemented by GPUs, ASICs, and FPGAs.[107] However, inference workloads are also fragmented between processing in both edge and cloud environments.

Cloud inference leverages the AI chips mentioned earlier and is provided as a service by CSPs. The market for this is reportedly large and fragmented due to the varied nature of the chips being used; however, Nvidia is likely to have a significant share here as well since GPUs can be easily programmed to move from training to inference workloads. However, even with a supply shortage for GPUs, CSPs are deploying their own proprietary chips (Google's TPU, for ex.), and plentifully available CPUs to run inference workloads. Therefore, this market segment is anticipated to be highly competitive[108].

Inference at the edge leverages chips or parts of SoCs in end-user devices like smartphones, cameras, and cars. While devices like smartphones tend to have SoCs built on leading-edge chips, most other devices utilise chips that can be made on mature fabrication processes. They are designed to use low power and to be cheap.[109] This market will also likely remain highly fragmented,

with Qualcomm, Intel and AMD continuing to provide chips for certain device categories like phones and laptops, but a wide range of companies exist to cater to various kinds of devices.[110]

Google's TPU is an example of a custom-designed accelerator chip that is optimised for both training and inference workloads.[111]

Characteristics of AI Chips Supply Chain

The supply chain for GPUs, as well as other AI accelerator chips and High Bandwidth Memory, is controlled by a select few countries that are dominant in the semiconductor global value chain (GVC), like Taiwan, South Korea, and the USA.[112] The technology, human capital, raw material and other inputs required to manufacture these highly specialised and intricate chips are spread out over these countries across a handful of companies at each stage of the value chain. Due to the hyper-globalized and specialised nature of the GVC, any disruption in the supply can severely affect the availability of chips for the end consumers. This dependency on a foreign entity-controlled supply chain can also become a strategic vulnerability, and foreign nations can potentially restrict access to any part of the supply chain.

The supply chain for other kinds of AI chips can be broadly divided into two categories: chips intended for training or inference workloads. The former,

like GPUs, are controlled by a handful of companies and countries. Integration into the GVC for these chips or creating a domestic ecosystem to manufacture homegrown alternatives will take immense financial expenditure over a long period to bear fruit.[113] On the other hand, inference chips can be more easily manufactured domestically, with architectures like RISC-V lending themselves well to AI inference workloads.[114] In this segment, policy intervention to ensure low entry barriers for domestic industry can happen via revamping trade policy to allow for a freer influx of input material and components required to manufacture such chips combined with industrial policies that enable the establishment of fabrication foundries on lower-cost mature processes which can tap into India's existing chip design talent pool.[115]

VIII. References

1 Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." Advances in Neural Information Processing Systems 25 (2012).

2 Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech Recognition with Deep Recurrent Neural Networks." arXiv.org, March 22, 2013. https://arxiv.org/abs/1303.5778.

3 Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." arXiv.org, June 12, 2017. https://arxiv.org/abs/1706.03762.

4 Benifei, Brando , and Ioan-Dragoş Tudorache. "Draft Compromise Amendments on the Draft Report Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," May 2023. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf.

5 Espinoza, Javier, Cristina Criddle, and Qianer Liu. "The Global Race to Set the Rules for AI." Financial Times, September 13, 2023. https://www.ft.com/content/59b9ef36-771f-4f91-89d1-ef89f4a2ec4e.

6 Nellis, Stephen. "Nvidia Turns to AI Cloud Rental to Spread New Technology." Reuters, March 21, 2023. https://www.reuters.com/technology/nvidia-set-reveal-new-ai-technologies-annual-conference-2023-03-21/.

7 Kelkar, Vijay, and Ajay Shah. In Service of the Republic: The Art and Science of Economic Policy. Penguin Random House India Private Limited, 2019.

8 OECD. "OECD Employment Outlook 2023: AI and Jobs, an Urgent Need to Act." OECD, July 11, 2023. https://www.oecd.org/employment-outlook/2023/.

9 Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. "On the Opportunities and Risks of Foundation Models." arXiv.org, August 16, 2021. https://arxiv.org/abs/2108.07258.

10 Andersen, Ross. "China's Artificial Intelligence Surveillance State Goes Global." The Atlantic, July 29, 2020. https://www.theatlantic.com/magazine/archive/2020/09/china-ai-surveillance/614197/.

11 Kotasthane, Pranay. "Takshashila Working Paper - High-Technology Geopolitics in the Post-Pandemic World — The Takshashila Institution." The Takshashila Institution, February 8, 2023. https://takshashila.org.in/research/takshashila-working-paper-high-technology-geopolitics-in-the-post-pandemic-world.

12 Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." arXiv.org, June 5, 2019. https://arxiv.org/abs/1906.02243.

13 AI Now Institute. "The Climate Costs of Big Tech." AI Now Institute, April 11, 2023. https://ainowinstitute.org/spotlight/climate.

14 Nokia. "MBiT Index 2023." Nokia, 2023. https://www.nokia.com/about-us/company/worldwide-presence/india/mbit-index-2023/.

15 Sahamati. "What Is Account Aggregator?" Sahamati, March 2, 2022. https://sahamati.org.in/what-is-account-aggregator/.

16 Huq, Aziz Z. "Who Owns Our Data?" Boston Review, October 26, 2021. https://www.bostonreview.net/articles/who-owns-our-data/.

17 Davenport, Thomas, and Ravi Kalakota. "The Potential for Artificial Intelligence in Healthcare." Future Healthcare Journal 6, no. 2 (June 2019): 94–98. https://doi.org/10.7861/futurehosp.6-2-94.

18 Shashidhara, Dr L S, Dr Vidya Mave, and Dr Joy Merwin. "Lessons from Pune: How Coordinated, Collaborative Data Driven Response Helped City Do Better in Second Surge." The Indian Express, May 15, 2021. https://indianexpress.com/article/cities/pune/opinion-pune-covid-second-wave-response-lessons-7316013/.

19 Sinha, Velu. "From Buzz to Reality: The Accelerating Pace of AI in India." Bain, June 28, 2022. https://www.bain.com/insights/from-buzz-to-reality-the-accelerating-pace-of-ai-in-india/.

20 Solaiman, Irene. "The Gradient of Generative AI Release: Methods and Considerations." arXiv.org, February 5, 2023. https://arxiv.org/abs/2302.04844.

21 Apar Gupta, Arjun Gargeyas, Bharath Reddy, Kailash Nadh, Nitin Pai, Pranay Kotasthane, Rushabh Mehta, Saurabh Chandra, and Venkatesh Hariharan. "An Open Tech Strategy for India — The Takshashila Institution." The Takshashila Institution, December 23, 2022. https://takshashila.org.in/research/an-open-tech-strategy-for-india.

22 Nagle, Frank. "Strengthening Digital Infrastructure: A Policy Agenda for Free and Open Source Software." Brookings, Brookings, 26 May 2022, https://www.brookings.edu/research/strengthening-digital-infrastructure-a-policy-agenda-for-free-and-open-source-software/.

23 Morgan, Forrest E., Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman. "Military Applications of AI Raise Ethical Concerns." RAND, January 1, 2020. https://www.rand.org/pubs/research_reports/RR3139-1.html.

24 NIST. "AI Risk Management Framework," July 12, 2021. https://www.nist.gov/itl/ai-risk-management-framework.

25 Lomas, Natasha. "Zoom Knots Itself a Legal Tangle over Use of Customer Data for Training AI Models." TechCrunch, August 9, 2023. https://techcrunch.com/2023/08/08/zoom-data-mining-for-ai-terms-gdpr-eprivacy/.

26 GitHub. "What Is GitHub Copilot?" GitHub. Accessed October 25, 2023. https://github.com/features/copilot.

27 Competition and Markets Authority. "AI Foundation Models: Initial Report." GOV.UK, September 18, 2023. https://www.gov.uk/government/publications/ai-foundation-models-initial-report.

28 Competition and Markets Authority. "AI Foundation Models: Initial Report." GOV.UK, September 18, 2023. https://www.gov.uk/government/publications/ai-foundation-models-initial-report.

29 Gao, Leo, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, et al. "The Pile: An 800GB Dataset of Diverse Text for Language Modeling." arXiv.org, December 31, 2020. https://arxiv.org/abs/2101.00027.

30 Zuboff, Shoshana. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Profile Books, 2019.

31 Lanier, Jaron. Who Owns the Future? 2014. Reprint, Simon and Schuster, 2014.

32 Arrieta-Ibarra, Imanol, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, and E. Glen Weyl. "Should We Treat Data as Labor? Moving Beyond 'Free.'" AEA Papers and Proceedings 108 (May 1, 2018): 38–42. https://doi.org/10.1257/pandp.20181003.

33 NITI Aayog. "Data Empowerment And Protection Architecture - Draft for Discussion," August 2020. https://www.niti.gov.in/sites/default/files/2020-09/DEPA-Book.pdf.

34 Office of the Law Revision Counsel. "17 U.S. Code § 107 - Limitations on Exclusive Rights: Fair Use." LII / Legal Information Institute. Accessed October 25, 2023. https://www.law.cornell.edu/uscode/text/17/107.

35 https://www.copyright.gov/fair-use/summaries/authorsguild-google-2dcir2015.pdf

36 Vincent, James. "The Lawsuit against Microsoft, GitHub and OpenAI That Could Change the Rules of AI Copyright." The Verge, November 8, 2022. https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data.

37 Vincent, James. "Getty Images Sues AI Art Generator Stable Diffusion in the US for Copyright Infringement." The Verge, February 6, 2023. https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion.

38 Creamer, Ella. "Authors File a Lawsuit against OpenAI for Unlawfully 'Ingesting' Their Books." The Guardian, July 5, 2023.

https://www.theguardian.com/books/2023/jul/05/authors-file-a-lawsuit-against-openai-for-unlawfully-ingesting-their-books.

39 Henderson, Peter, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. "Foundation Models and Fair Use." arXiv. Accessed November 18, 2023. https://arxiv.org/pdf/2303.15715.pdf.

40 pnp. "Computational Power and AI." AI Now Institute, September 27, 2023. https://ainowinstitute.org/publication/policy/compute-and-ai.

41 The White House. "FACT SHEET: Advancing Technology for Democracy." The White House, March 29, 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/03/29/fact-sheet-advancing-technology-for-democracy-at-home-and-abroad/.

42 CBRE India. "The Evolving Landscape of Data Centre in India." Accessed November 16, 2023. https://www.cbre.co.in/insights/articles/the-evolving-landscape-of-data-centre-in-india. .

43 Sarkar, Soumyadeep. "Google Unveils Pixel 8 and Pixel 8 Pro, Pre-Orders Start Today." The Tech Portal, October 5, 2023. https://thetechportal.com/2023/10/05/google-pixel-8-pixel-8-pro-launch-specs/.

44 Noone, Greg. "Is the Cloud Computing Market Anti-Competitive?" Tech Monitor, November 25, 2021. https://techmonitor.ai/technology/cloud/is-cloud-computing-market-anti-competitive-antitrust.

45 Advocates, Verus. "Cloud Computing – Legal and Policy Perspectives." Career Intelligence for Lawyers, Law Students, July 24, 2017. https://www.legallyindia.com/home/cloud-computing-legal-and-policy-perspectives-20170724-8678.

46 Competition, and Markets Authority. "AI Foundation Models: Initial Report." GOV.UK, September 18, 2023. https://www.gov.uk/government/publications/ai-foundation-models-initial-report.

47 Vivek, Sonu. "What Does India's 25,000 GPU Proposal Mean for AI Startups?" TICE News, October 6, 2023. https://www.tice.news/tice-trending/indian-ai-startups-gpu-innovation-trending-1512829.

48 Telecom Regulatory Authority of India. "Recommendations on Consumer Protection in the Telecom Sector." September 14, 2020. PDF. https://www.trai.gov.in/sites/default/files/Recommendations_CS_14092020.pdf.

49 The Digital Personal Data Protection Act, 2023 (No.22 of 2023) (n.d.). https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf.

50 Vengattil, Munsif, and Dhwani Pandya. "Nvidia Strikes Deals with Reliance, Tata in Deepening India AI Bet." Reuters, September 8, 2023. https://www.reuters.com/technology/nvidia-reliance-partners-develop-ai-apps-india-2023-09-08/.

51 Competition, and Markets Authority. "AI Foundation Models: Initial Report." GOV.UK, September 18, 2023. https://www.gov.uk/government/publications/ai-foundation-models-initial-report.

52 Patel, Dylan. "How Nvidia's CUDA Monopoly In Machine Learning Is Breaking - OpenAI Triton And PyTorch 2.0." SemiAnalysis, January 16, 2023. https://www.semianalysis.com/p/nvidiaopenaitritonpytorch.

[53] Shilov, Anton. "Nvidia Bans Using Translation Layers for CUDA Software — Previously the Prohibition Was Only Listed in the Online…." Tom's Hardware, March 4, 2024. https://www.tomshardware.com/pc-components/gpus/nvidia-bans-using-translation-layers-for-cuda-software-to-run-on-other-chips-new-restriction-apparently-targets-zluda-and-some-chinese-gpu-makers.

[54] ETtech. "India to Become AI Use Case Capital of the World: Nandan Nilekani." Economic Times, May 7, 2024. https://economictimes.indiatimes.com/tech/technology/india-to-become-ai-use-case-capital-of-the-world-nandan-nilekani/articleshow/109926942.cms?from=mdr.

[55] Mohanty, Amlan, and Adarsh Ranjan. "India's Compute Conundrum." Carnegie Endowment for International Peace, February 2024. https://carnegieendowment.org/posts/2024/02/indias-compute-conundrum?lang=en. (Accessed 13 June 2024)

[56] Bhatnagar, Rishabh. "UPI Moment For Compute, An Indian Fintech LLM, Sarvam's First Demo: Key Takeaways From Adbhut India." NDTV Profit, May 8, 2024. https://www.ndtvprofit.com/technology/upi-moment-for-compute-an-indian-fintech-llm-sarvams-first-demo-key-takeaways-from-adbhut-india.

[57] ONDC | Open Network for Digital Commerce. "ONDC." Accessed June 13, 2024. https://ondc.org/.

[58] Bhatnagar, Rishabh. "UPI Moment For Compute, An Indian Fintech LLM, Sarvam's First Demo: Key Takeaways From Adbhut India." NDTV Profit, May 8, 2024. https://www.ndtvprofit.com/technology/upi-moment-for-compute-an-indian-fintech-llm-sarvams-first-demo-key-takeaways-from-adbhut-india.

[59] Manur, Anupam. "Flawed Regulation Can Undermine the Digital Payment Ecosystem." Hindustan Times. https://www.hindustantimes.com/analysis/flawed-regulation-can-undermine-the-digital-payment-ecosystem/story-jVikN3SWxYYRKGZZiO5loL.html. (Accessed 13 June 2024)

60 The MIT Press Reader. "The Staggering Ecological Impacts of Computation and the Cloud." The MIT Press Reader, February 14, 2022. https://thereader.mitpress.mit.edu/the-staggering-ecological-impacts-of-computation-and-the-cloud/.

61 BairesDev Blog: Insights on Software Development & Tech Talent. "No Longer Hidden in the Cloud: The Push for Sustainable Technology," August 11, 2023. https://www.bairesdev.com/blog/no-hidden-cloud-push-sustainable-technology/.

62 Kumar, Kalyan. "Towards a Sustainable Transition to the Green Cloud." CMSWire.Com, October 19, 2021. https://www.cmswire.com/information-management/towards-a-sustainable-transition-to-the-green-cloud/.

63 BairesDev Blog: Insights on Software Development & Tech Talent. "No Longer Hidden in the Cloud: The Push for Sustainable Technology," August 11, 2023. https://www.bairesdev.com/blog/no-hidden-cloud-push-sustainable-technology/.

64 Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." arXiv.org, June 12, 2017. https://arxiv.org/abs/1706.03762.

65 Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "LoRA: Low-Rank Adaptation of Large Language Models." arXiv.org, June 17, 2021. https://arxiv.org/abs/2106.09685.

66 "AI Index Report 2023 – Artificial Intelligence Index." Accessed November 18, 2023. https://aiindex.stanford.edu/report/.

67 Chahal, Husanjot, Sara Abdulla, Jonathan Murdick, and Ilya Rahkovsky. "Mapping India's AI Potential." Center for Security and Emerging Technology, March 2021. https://cset.georgetown.edu/wp-content/uploads/CSET-Mapping-Indias-AI-Potential-Report.pdf.

68 Techtracker ASPI. "Artificial Intelligence Algorithms and Hardware Accelerators." Accessed November 18, 2023. https://techtracker.aspi.org.au/tech/artificial-intelligence-algorithms-and-hardware-accelerators/?c1=in&colours=true.

69 Lab, Scimago. "Scimago Country Rankings." International Science Ranking. Accessed November 18, 2023. https://www.scimagojr.com/countryrank.php?category=1702&order=h&ord=desc.

70 Department of Science and Technology, Government of India. 2023. R&D Statistics at a Glance, 2022-23. New Delhi: Department of Science and Technology, Government of India. https://dst.gov.in/sites/default/files/R%26D%20Statistics%20at%20a%20Glance%2C%202022-23.pdf.

71 Sinha, Velu. "From Buzz to Reality: The Accelerating Pace of AI in India." Bain, June 28, 2022. https://www.bain.com/insights/from-buzz-to-reality-the-accelerating-pace-of-ai-in-india/.

72 MacroPolo. "The Global AI Talent Tracker," June 9, 2020. https://macropolo.org/digital-projects/the-global-ai-talent-tracker/.

73 Edgerton, Anna and Bass, Dina."Microsoft Calls for New U.S. Agency and Licensing for AI Tools." Bloomberg, May 25, 2023. https://www.bloomberg.com/news/articles/2023-05-25/microsoft-calls-for-new-us-agency-and-licensing-for-ai-tools

74 Future of Life Institute. "Pause Giant AI Experiments: An Open Letter," March 22, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

75 Kapoor, Sayash, and Arvind Narayanan. "A Misleading Open Letter about Sci-Fi AI Dangers Ignores the Real Risks." AI Snake Oil, March 29, 2023. https://www.aisnakeoil.com/p/a-misleading-open-letter-about-sci.

76 Liesenfeld, Andreas, Alianda Lopez, and Mark Dingemanse. "Opening up ChatGPT: LLM Openness Leaderboard." Opening up ChatGPT, July 2023. https://opening-up-chatgpt.github.io/.

77 Solaiman, Irene. "The Gradient of Generative AI Release: Methods and Considerations." arXiv.org, February 5, 2023. https://arxiv.org/abs/2302.04844.

78 Big Science. "BLOOM." Accessed November 18, 2023. https://bigscience.huggingface.co/blog/bloom.

79 Patel, Dylan, and Afzal Ahmad. "Google 'We Have No Moat, And Neither Does OpenAI.'" SemiAnalysis, May 4, 2023. https://www.semianalysis.com/p/google-we-have-no-moat-and-neither.

80 Foster, Kelsey. "PyTorch vs TensorFlow: Who Has More Pre-Trained Deep Learning Models?" HackerNoon, March 20, 2022. https://hackernoon.com/pytorch-vs-tensorflow-who-has-more-pre-trained-deep-learning-models.

81 Meta. "Llama 2." Meta AI. Accessed November 18, 2023. https://ai.meta.com/llama/.

82 Narang, Sharan, and Aakanksha Chowdhery. "Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance," April 4, 2022. https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html.

83 Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots." In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: ACM, 2021. http://dx.doi.org/10.1145/3442188.3445922.

84 Knight, Will. "OpenAI's CEO Says the Age of Giant AI Models Is Already Over." WIRED, April 17, 2023. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

85 Wheatley, Mike. "Microsoft Hedges Its Bets, Seeking More Cost-Effective AI Models." SiliconANGLE, September 26, 2023. https://siliconangle.com/2023/09/26/microsoft-hedges-bets-seeking-cost-effective-ai-models/.

86 NIST. "AI Risk Management Framework," July 12, 2021. https://www.nist.gov/itl/ai-risk-management-framework.

87 Kotasthane, Pranay. "Takshashila Working Paper - High-Technology Geopolitics in the Post-Pandemic World — The Takshashila Institution." The Takshashila Institution, February 8, 2023. https://takshashila.org.in/research/takshashila-working-paper-high-technology-geopolitics-in-the-post-pandemic-world.

88 ibid.

89 McKinsey & Company. "Artificial Intelligence Hardware: New Opportunities for Semiconductor Companies." Accessed November 16, 2023. PDF. https://www.mckinsey.com/~/media/McKinsey/Industries/Semiconductors/Our%20Insights/Artificial%20intelligence%20hardware%20New%20opportunities%20for%20semiconductor%20companies/Artificial-intelligence-hardware.ashx.

90 Uz, Fidan Boylu. "GPUs vs CPUs for Deployment of Deep Learning Models." Microsoft Azure Blog (blog), September 11, 2018. https://azure.microsoft.com/en-us/blog/gpus-vs-cpus-for-deployment-of-deep-learning-models/.

91 Leswing, Kif. "Meet the $10,000 Nvidia Chip Powering the Race for A.I." CNBC, February 23, 2023. https://www.cnbc.com/2023/02/23/nvidias-a100-is-the-10000-chip-powering-the-race-for-ai-.html.

92 Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, et al. "Training Compute-Optimal Large Language Models." arXiv.org, March 29, 2022. https://arxiv.org/abs/2203.15556.

93 Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, et al. "Training Compute-Optimal Large Language Models." arXiv.org, March 29, 2022. https://arxiv.org/abs/2203.15556.

94 Competition, and Markets Authority. "AI Foundation Models: Initial Report." GOV.UK, September 18, 2023. https://www.gov.uk/government/publications/ai-foundation-models-initial-report.

95 Knight, Will. "OpenAI's CEO Says the Age of Giant AI Models Is Already Over." WIRED, April 17, 2023. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

96 Mok, Aaron. "ChatGPT Could Cost over $700,000 per Day to Operate. Microsoft Is Reportedly Trying to Make It Cheaper." Business Insider India, April 20, 2023. https://www.businessinsider.in/tech/news/chatgpt-could-cost-over-700000-per-day-to-operate-microsoft-is-reportedly-trying-to-make-it-cheaper-/articleshow/99637548.cms.

97 Patrizio, Andy. "Nvidia Still Crushing the Data Center Market." Network World, December 22, 2022. https://www.networkworld.com/article/3684174/nvidia-still-crushing-the-data-center-market.html.

98 Heller, Martin. "What Is CUDA? Parallel Programming for GPUs." InfoWorld, September 16, 2022. https://www.infoworld.com/article/3299703/what-is-cuda-parallel-programming-for-gpus.html.

99 Patrizio, Andy. "Nvidia Still Crushing the Data Center Market." Network World, December 22, 2022. https://www.networkworld.com/article/3684174/nvidia-still-crushing-the-data-center-market.html.

100 Patrizio, Andy. "Nvidia Still Crushing the Data Center Market." Network World, December 22, 2022. https://www.networkworld.com/article/3684174/nvidia-still-crushing-the-data-center-market.html.

101 VK, Anirudh. "NVIDIA Emerges As Clear Winner in AI Showdown." Analytics India Magazine, January 24, 2023. https://analyticsindiamag.com/nvidia-emerges-as-clear-winner-in-ai-showdown/.

102 Ware, Ana. “Infrastructure Requirements for AI Inference vs. Training.” HPCwire, June 13, 2022. https://www.hpcwire.com/2022/06/13/infrastructure-requirements-for-ai-inference-vs-training/.

103 Ware, Ana. “Infrastructure Requirements for AI Inference vs. Training.” HPCwire, June 13, 2022. https://www.hpcwire.com/2022/06/13/infrastructure-requirements-for-ai-inference-vs-training/.

104 Gargeyas, Arjun, Samparna Tripathy, and Anup Rajput. “Takshashila Discussion SlideDoc - An AI Hardware Ecosystem in India: A SWOT Analysis — The Takshashila Institution.” The Takshashila Institution, September 26, 2022.

105 Goldberg, Jonathan. “The AI Chip Market Landscape: Choose Your Battles Carefully.” TechSpot, May 30, 2023. https://www.techspot.com/news/98879-ai-chip-market-landscape-choose-battles-carefully.html.

106 Ware, Ana. “Infrastructure Requirements for AI Inference vs. Training.” HPCwire, June 13, 2022. https://www.hpcwire.com/2022/06/13/infrastructure-requirements-for-ai-inference-vs-training/.

107 IBM. "Why Your AI Infrastructure Needs Both Training and Inference Platforms." October 17, 2019. PDF. https://www-2000.ibm.com/partnerworld/pdfs/why-your-ai-infrastructure-needs-both-training-and-inference-platforms-10-17-19_94028894USEN.pdf.

108 Goldberg, Jonathan. “The AI Chip Market Landscape: Choose Your Battles Carefully.” TechSpot, May 30, 2023. https://www.techspot.com/news/98879-ai-chip-market-landscape-choose-battles-carefully.html.

109 Gargeyas, Arjun, Samparna Tripathy, and Anup Rajput. "Takshashila Discussion SlideDoc - An AI Hardware Ecosystem in India: A SWOT Analysis — The Takshashila Institution." The Takshashila Institution, September 26, 2022. https://takshashila.org.in/research/an-ai-hardware-ecosystem-in-india-a-swot-analysis.

110 Goldberg, Jonathan. "The AI Chip Market Landscape: Choose Your Battles Carefully." TechSpot, May 30, 2023. https://www.techspot.com/news/98879-ai-chip-market-landscape-choose-battles-carefully.html.

111 Google Cloud. "Tensor Processing Units (TPUs) ." Accessed November 16, 2023. https://cloud.google.com/tpu#:~:text=Google%20Cloud%20TPUs%20are%20custom,inference%20of%20large%20AI%20models.

112 INDIAai. "Semiconductors - at the Core of Artificial Intelligence Technologies." Accessed November 16, 2023. https://indiaai.gov.in/article/semiconductors-at-the-core-of-artificial-intelligence-technologies.

113 Ghai, Palak. "Why Only Few Companies Produce Semiconductor Chips?" StartupTalky, July 4, 2022. https://startuptalky.com/why-few-companies-produce-semiconductor/.

114 https://takshashila.org.in/research/an-ai-hardware-ecosystem-in-india-a-swot-analysis

115 Chintapali, Rohit. "Design-Focused Schemes More Suited For India's Semiconductor Aspirations." BW Businessworld. Accessed November 16, 2023. https://www.businessworld.in/article/Design-focused-Schemes-More-Suited-For-India-s-Semiconductor-Aspirations/24-04-2023-473973/.

The Takshashila Institution is an independent centre for research and education in public policy. It is a non-partisan, non-profit organisation that advocates the values of freedom, openness, tolerance, pluralism, and responsible citizenship. It seeks to transform India through better public policies, bridging the governance gap by developing better public servants, civil society leaders, professionals, and informed citizens.

Takshashila creates change by connecting good people, to good ideas and good networks. It produces independent policy research in a number of areas of governance, it grooms civic leaders through its online education programmes and engages in public discourse through its publications and digital media.