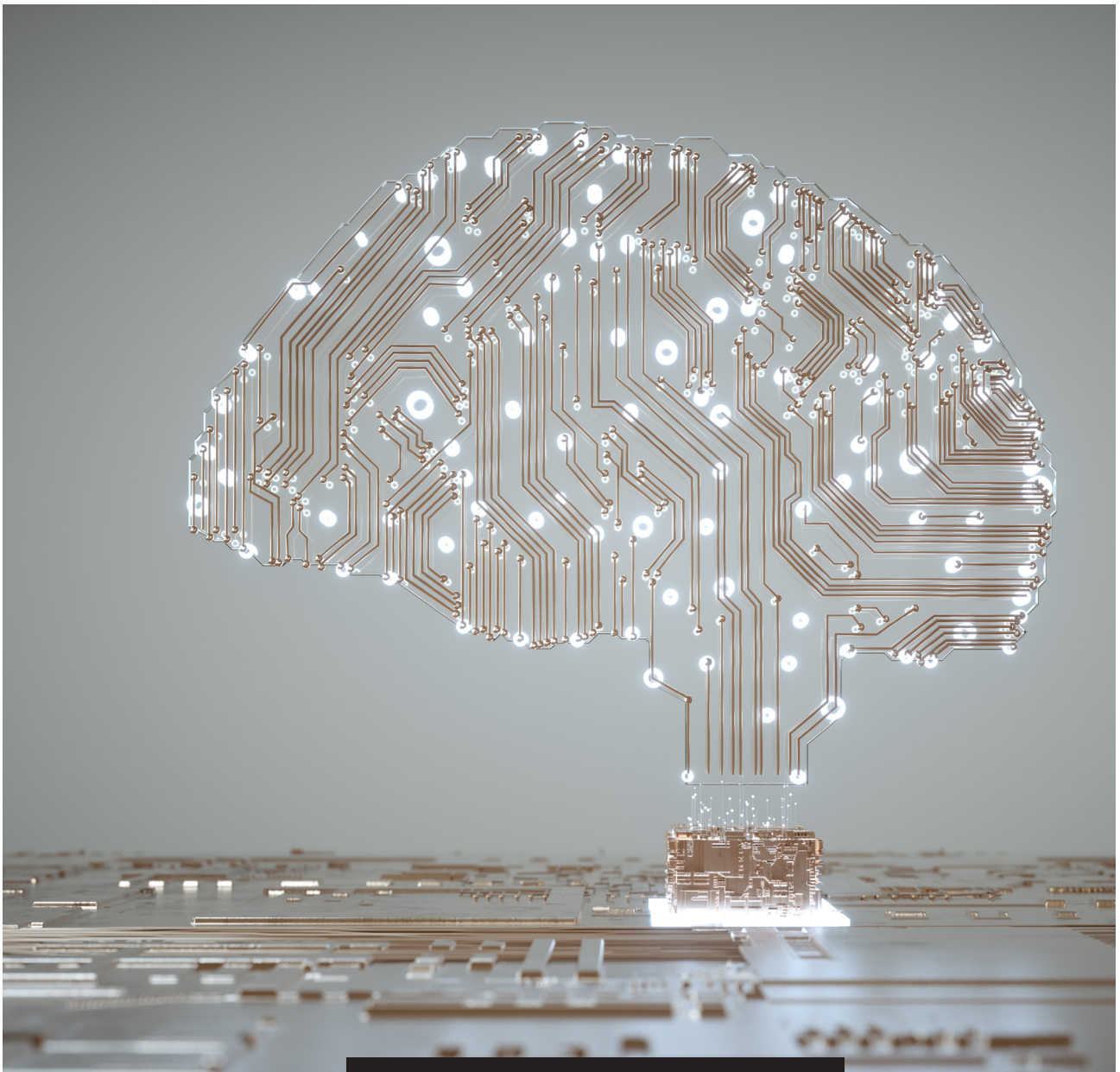


SPECIAL **REPORT** no. 242

Artificial Intelligence: Hardware, Interface, and Applications

Prateek Tripathi



DECEMBER 2024

Abstract

Artificial Intelligence (AI) has led to a fundamental shift in the human-machine interface, with massive implications for the future. AI today has known applications across multiple domains—including agriculture, defence, healthcare, finance, manufacturing, and nuclear energy—and the potential appears limitless. Justifiably then, the discourse surrounding AI

is becoming increasingly more vibrant. Yet, the inner workings of AI are often shrouded in mystery, with the term itself often misinterpreted and misused. Before any meaningful discussion on AI can proceed, it is important to understand what AI means and how it works, as well as the hardware requirements for training AI models.

Attribution: Prateek Tripathi, “Artificial Intelligence: Hardware, Interface, and Applications,” *ORF Special Report No. 242*, December 2024, Observer Research Foundation.

AI: An Overview

Types of AI

‘Artificial intelligence’ is a blanket term, often used to represent a multitude of ideas. The terms ‘machine learning’ (ML), ‘deep learning’ and ‘neural networks’ are all used interchangeably for AI.¹ The simplest way to understand these terms is to envision them as a series of evolving systems, each encompassing the next. AI is the overarching system, with ML a subset; deep learning, in turn, is a subfield of ML; and neural networks make up the backbone of deep learning algorithms.² It is the depth of a neural network, or the number

of node layers, that distinguishes a single neural network from a deep learning algorithm, which must have more than three layers.³

Neural Networks

Neural networks are modelled to mimic the way neurons function in a human brain.⁴ They consist of huge volumes of densely interconnected nodes organised into layers—an input layer, one or more hidden layers, and an output layer.⁵ An individual node may be connected to several nodes from the layers above and below it, receiving data from the former and sending data to the latter. These layers are ‘feed-forward,’ meaning the flow of data occurs only in one direction.⁶

A node assigns a number known as a ‘weight’ to each of its incoming connections.⁷ When the network is active, each time a node receives a different data item—holding a different number—over any of its connections, it multiplies it by the associated weight of the connection. It then adds the products together, to yield a single number. If that number is below a threshold value, the node does not pass any data to the next layer. If the number exceeds the threshold value, the node ‘fires,’ meaning it sends the number, which is the sum of all the weighted inputs, along all its outgoing connections.

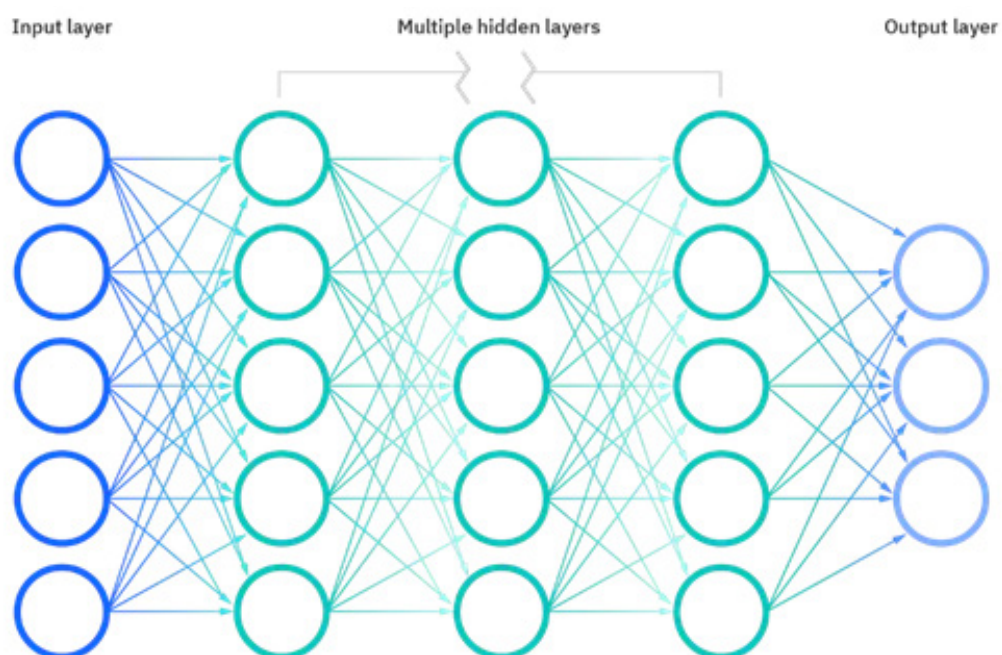
When a neural network is being trained, all its weights and thresholds are initially set to random values. Training data is fed to the bottom or input layer. It passes through the succeeding layers, getting multiplied and added together in complex ways, until it finally arrives, radically transformed, at the output layer. The weights and thresholds are continually adjusted until training data with the same labels consistently yields similar outputs.⁸ Google’s search algorithm is an example of a neural network.⁹

Deep Learning

The ‘deep’ in deep learning refers to the depth of layers in a neural network—a neural network consisting of more than three node layers, including the input and output layers, is referred to as a ‘deep learning neural network’.¹⁰ Essentially a subset of ML, deep learning enhances optimisation and refines accuracy over single-layer neural networks, thanks to its additional ‘hidden layers.’¹¹ Deep learning combines forward propagation, the progression of computations through its network, with backpropagation, the use of certain algorithms to calculate errors in predictions and subsequently adjust the weights and biases of the nodes by moving backwards through the layers to train the model.¹² Together, these processes allow a neural network to make predictions and error-corrections, gradually increasing the accuracy of the model over time.

Different kinds of deep learning neural networks exist to address specific problems. For instance, convolutional neural networks are used in image classification and object detection applications,¹³ while recurrent neural networks find application in natural language and speech recognition.¹⁴

Figure 1. Deep Neural Networks



Source: IBM¹⁵

Table 1. Deep Learning Start-ups in India

Start-up	Description	Total Funding Raised
1. Qure.ai ¹⁶	Leverages deep learning to provide automated interpretation of radiology exams like X-rays and ultrasounds.	US\$123 million
2. Whiterabbit.ai ¹⁷	Uses deep-learning based software to detect breast cancer.	US\$49 million
3. Nanonets ¹⁸	Uses deep learning models to automate key business processes like calculating and structuring accounts payable, order processing and insurance underwriting.	US\$42 million
4. ParallelDots ¹⁹	Uses deep learning algorithms for image recognition software which helps Consumer Packaged Goods manufacturers in optimising in-store exposure and maximising sales.	US\$6.5 million

Source: Tracxn²⁰

Machine Learning

The use of ML is visible in popular applications like Siri as well as new technology like self-driving cars. While ML and deep learning operate in a similar fashion, there are subtle differences between the two. Deep learning can process unstructured data in its raw form, and automatically determine the features that distinguish one category of data from another.²¹ It can thus function independently of human intervention and can operate on larger datasets. This is why it is also referred to as ‘scalable ML.’

ML, on the other hand, is more dependent on human intervention to learn and requires human experts to determine the set of features that differentiate data inputs from each other, a process known as “feature extraction.”²² It thus requires more structured data and usually operates on smaller datasets as compared to deep learning.²³

There are four main types of ML:²⁴

1. *Supervised ML*: Supervised learning employs labeled datasets to train algorithms to classify data or predict outcomes accurately. The model adjusts its weights based on the data fed into it.

2. *Unsupervised ML*: Unsupervised learning uses ML algorithms to analyse and cluster unlabeled datasets. It is used to detect hidden data patterns or groupings without the need for human intervention. However, it still requires human intervention to validate output variables.
3. *Semi-supervised learning*: This method uses a smaller labeled dataset to guide classification and feature extraction from a larger, unlabeled dataset. It is useful when there is a lack of sufficient labeled data or it is too costly to label data.
4. *Reinforcement ML*: Reinforcement ML is similar to supervised learning, the difference being that in this case, the algorithm is not trained using sample data, rather the model proceeds via trial and error. A sequence of successful outcomes is reinforced to give the best model for a given problem.

The use of ML thus requires large training datasets that are accurate and unbiased.²⁵ Gathering sufficient data and having a system robust enough to run it can be a drain on resources. ML can also be prone to error, depending on the input. With too small a sample, the system could produce a perfectly logical algorithm that is completely wrong or misleading.

Table 2. ML Start-ups in India

Start-up	Description	Total Funding Raised
1. Uniphore ²⁶	Uses ML models to aid business operations by enhancing customer and employee experiences through the use of multi-modal AI and data platforms.	US\$657 million
2. Eightfold ²⁷	Uses ML algorithms to provide insights and assistance to job seekers and employers.	US\$410 million
3. SigTuple ²⁸	Develops ML and AI-based diagnostic solutions for the medical industry.	US\$54.7 million
4. Haptik ²⁹	Uses ML and natural language processing (NLP) to develop conversational AI platforms.	US\$12.2 million

Source: Tracxn³⁰

What is Generative AI?

The rapid development in neural networks, deep learning and ML led to the emergence of a novel kind of AI, called ‘Generative AI’. The foundation for Generative AI was laid by the inception of natural language processing (NLP), which is an ML model that give computers the ability to interpret, manipulate, and comprehend human language.³¹ Around 2010, AI researchers working

on natural language translation discovered that AI models trained on vast amounts of text produced much better results than those using top-down grammatical tools. This led to the development of so-called large language models (LLMs), which were trained explicitly on large amounts of text data for NLP tasks, and contained a huge number of parameters, usually exceeding 100 million.³² LLMs facilitated the processing and generation of natural language text for diverse tasks.

However, since each model had its strengths and weaknesses, this necessitated developing multiple models, depending on the specific NLP task and the characteristics of the data being analysed. Creating and deploying each new LLM required a considerable amount of time and resources. Each new application required a large, well-labelled dataset, depending on the specific objective that needed to be achieved.³³ In case a dataset was unavailable, finding and labelling appropriate images, text or graphs for it required hundreds to thousands of hours. Additionally, training one LLM had roughly the same carbon footprint as running five cars over their lifetime.³⁴

This created the necessity for a new generation of AI models to replace the task-specific models which had dominated the AI landscape up to that point, and led to models trained on a broad set of unlabelled data that could be applied for a performing a wide variety of tasks, with minimal fine-tuning. The result was the ‘foundation models’, a term popularised by the Stanford Institute for Human-Centered Artificial Intelligence.³⁵ Early examples of these models included GPT-3, BERT, or DALL-E 2. Using a short input prompt, such systems could generate

an entire essay, or a complex image, based on the given parameters, even if they were not specifically trained on how to execute that exact argument or generate an image in that way.

Generative AI emerged as a consequence of a new neural network architecture called a ‘transformer’.³⁶ Most assign credit to OpenAI’s 2018 invention of the Generative Pre-trained Transformer (GPT), a new type of LLM with significant improvements in natural language understanding. Yet, at the core of a GPT is the earlier innovation of the transformer architecture that enabled the parallel processing required to capture long-term context around natural language inputs.

Combining transformer architecture with unsupervised or semi-supervised learning, large foundation models emerged. Capable of handling multiple data modalities, these outperformed existing benchmarks and could apply information learnt in one situation to another.³⁷ This led to the creation of generative AI, which, as the name suggests, is capable of generating text, images, music, video or code, by interpreting and manipulating pre-existing data.

Hardware Requirements for AI

Training AI algorithms is an arduous task that requires a considerable amount of hardware. Some of the essential hardware requirements are outlined in the following points.

1. Graphical Processing Units (GPUs)

It is the recent resurgence in neural networks that is, in fact, the driving force behind the tremendous advancements in AI in the past few years. This improvement in neural networks, in turn, owes to the gaming industry and its use of GPUs.

GPUs, which pack thousands of processing cores on a single chip, and are required to run increasingly evolving video games, function in a

way very similar to neural networks,³⁸ thus offering significant performance improvements over CPUs for both deep learning and ML training.

Training is the most computationally intensive task in preparing any ML model, involving multiple steps, including pre-processing the input data, training the model, storing the trained model and deploying the model.³⁹ Training models faster thus requires the ability to perform multiple operations simultaneously instead of one at a time. This is where GPUs have an edge over CPUs, since they contain thousands of cores designed to compute with almost 100 percent efficiency, offsetting the faster speed provided by CPUs.⁴⁰

Training large AI models thus typically requires the use of GPU clusters. Meta, for instance, has created two new data centre-scale clusters designed to support larger and more complex AI models like its Llama 3.⁴¹ The clusters each contain 24,576 Nvidia H100 GPUs, while Meta's original clusters contained about 16,000 Nvidia A100 GPUs. By the end of 2024, the company is aiming to grow its infrastructure build-out to include 350,000 Nvidia H100s as part of a portfolio that will feature compute power equivalent to almost 600,000 H100s.⁴²

2. Network Connectivity

Training AI models is a network-intensive task that requires powerful and efficient networking components and setups. Concurrently, the data centres and cloud networks designed for them are also unique and more specialised since there is a need to integrate GPUs and data processing units (DPUs) into the computing and storage hardware to accelerate AI training and workloads.⁴³ Low latency along with high bandwidth and stability are critical requirements for training AI models. While traditional network architectures are designed to process numerous but small workloads, a large proportion of the

time processing AI workloads occurs on the network.⁴⁴ Additionally, training large AI models requires clusters comprising thousands of GPUs. This poses a challenge since it necessitates that the networking and computational capabilities be highly matched. Typical cloud computing servers possess a bandwidth of up to 100 giga bits per second, whereas the bandwidth requirement for AI platforms can reach the tera bits per second range, which represents over a 100-fold jump in bandwidth capacity.⁴⁵ This requires specialised technologies like remote direct memory access and network architectures like InfiniBand, which are more cost-intensive than traditional networks.

3. Memory and Storage

In addition to processor requirements, memory and storage are other key considerations for the AI/ML pipeline. To train or operate a ML model, programs require data and code to be stored in local memory to be executed by the processor.⁴⁶ Some models, like decision trees, may be trained with less memory because the algorithms are smaller. Others, like deep neural networks, may require faster local memory because the algorithms are larger. High-bandwidth memory and solid-state drives are therefore becoming increasingly important for training and running these models efficiently.⁴⁷

Though cloud storage in a distributed file system typically removes any storage limitations historically imposed by the local hard disk size,⁴⁸ many real-world AI/ML use cases involve complex, multi-step pipelines. Each step could require different libraries and runtimes and may need to execute on specialised hardware profiles.⁴⁹ Therefore, local disk storage is still an important factor despite the presence of cloud storage.

4. Power

AI development can essentially be broken down into two phases. First is the training phase, which has already been discussed. Second is an inference phase, where the model is put into live operation and fed prompts so it can produce original responses. Both these phases are energy-intensive, though the exact division of energy requirement is not fully understood yet.⁵⁰ With the Google search engine, for instance, 60 percent energy went into the inference stage and 40 percent into training.⁵¹ The situation changed with ChatGPT, however, since it required very low energy consumption during the training phase in comparison with

actually applying the model.⁵² Energy consumption scales up with the size of the datasets being used, among other factors.

5. ASICs and FPGAs

Unlike GPUs, application-specific integrated circuits (ASICs) and field-programmable gate arrays (FPGAs) are low-power and cost-effective chips designed specifically for AI applications.⁵³

ASICs are designed explicitly for particular applications or tasks and offer a performance advantage for specific AI workloads.⁵⁴ However, they do suffer from a lack of flexibility. Tensor Processing Units, custom designed by Google for TensorFlow, one of the most popular deep learning frameworks, are an example of an ASIC.⁵⁵

FPGAs are integrated circuits or chips with a programmable hardware fabric.⁵⁶ Unlike GPUs or ASICs, the circuitry inside an FPGA chip is not hard-etched—it can be reprogrammed as needed. This capability makes FPGAs an excellent alternative to ASICs, which require a long development time and a significant investment to design and fabricate.⁵⁷ They also offer increased flexibility over ASICs.

While FPGAs are mostly used to apply trained AI algorithms to real-world data inputs (or inference), ASICs can be designed for either training or inference.⁵⁸

6. AI supercomputers

Training unique AI foundation models is an expensive and resource-intensive task, necessitating virtual supercomputers. These are called AI supercomputers, with Nvidia being one of the first companies to describe its systems as such.⁵⁹ The main functional difference between supercomputers and AI supercomputers often lies in the math format they use for computation.

Traditional supercomputers use double-precision (64-bit) floating-point performance, while AI supercomputers focus on lower-precision math, which may scale down to eight-bit floating point performance.⁶⁰ These are used for model training since neural networks do not require higher precision.

Additionally, AI supercomputers are often locked down, on-premise machines with a very specific bare-metal design and unique networking backbones.⁶¹ They are usually massive, with thousands of CPUs and GPUs, and thus also require fast and reliable network connectivity. For example, Microsoft Azure's OpenAI supercomputer is 'purpose-built,' with a specialised 400 gigabits per second network connectivity for each GPU server.⁶²

AI Interface

In addition to the types of AI and the accompanying hardware requirements, another important aspect with regard to AI is the profound impact it has had on human-to-machine interface. A user interface (UI) is the means we use to interact with any digital device, be it a computer, smartphone or even an ATM.⁶³ The most common UI is the graphical user interface (GUI), which was initially popularised by Macintosh and Microsoft Windows, and is now prevalent in virtually all computing devices.⁶⁴ In a GUI, the user interacts with the device via icons, menus or other visual representations, through a

point-and-click device like a mouse, or fingers in the case of a touch screen. Any novel UI is based on unlocking a new abstraction layer to hide the working details of a subsystem; generalizing details allows complex operating systems to appear simpler and more intuitive.⁶⁵

The emergence of AI, particularly generative AI, has led to a fundamental shift in digital interface. Up to this point, all UIs were based on the idea of the user describing the exact process required to execute the intended outcome.

AI has fundamentally altered this procedure by introducing ‘intent-based’ UI, wherein the user simply states the desired outcome, relinquishing control over the process to the AI.⁶⁶ The interfaces of most generative AI applications now revolve around an input field in which a user can type anything to prompt the system. This has reduced the interaction between the user and the computer dramatically, requiring re-engagement only if the results are unsatisfactory.⁶⁷ Similar to how GUIs superseded command lines, conversational interfaces like AI chatbots are now replacing GUIs. GPTs have allowed conversational interfaces to organise unstructured datasets to create responses with human-like (or greater) intelligence.⁶⁸

This new UI paradigm introduced by generative AI, however, comes with its own caveats, the primary one being deep-rooted usability problems. The core issue with conversational interfaces is that they offload technical work to non-technical users, making the ideal output only attainable through learned commands.⁶⁹ This has led to the emergence of ‘prompt engineering,’ the practice of tailoring inputs to best communicate with generative AI.⁷⁰ The chat-based interaction

style employed by generative AI tools also suffers from requiring users to write out their inputs as prose text. This inherently requires the user base to be literate and articulate enough to get good results from AI chatbots, which severely limits their accessibility.⁷¹

Whether this intent-based UI offered by generative AI tools can truly offer a viable replacement to the existing UI paradigm, remains an open question. Visual information is often easier to understand, process and interact with. For example, filling out an application form with an AI chatbot is far more tedious than employing a GUI. It is likely that future AI systems will possess a hybrid user interface that combines elements of intent-based and command-based interfaces while still retaining many GUI elements.⁷² Zero-command interfaces, such as retinal scanners and toll transponders, provide another possible UI paradigm for the future.⁷³ In this case, AI would be able to produce the desired outcome without the user having to enter any command whatsoever. This could also eventually lead to brain-to-computer interfaces (BCI), where AI would read brain scans to decipher thoughts.⁷⁴

The Applications of AI

The growth in the applications of AI in recent years has been remarkable. A 2023 study of 34 countries with a national AI strategy identified healthcare, transportation, information and communication technology, natural resources and energy, agriculture, and education, as the focal areas for the application of AI.⁷⁵ In the Indian context, NITI Aayog's National Strategy for Artificial Intelligence lists healthcare, agriculture, education, smart cities, and smart mobility as some of the key sectors for AI deployment.⁷⁶ In addition to these domains, AI also holds massive potential in helping nations achieve their sustainable development goals, in areas such as the adoption of clean energy and management of coastal zones. Though most of the countries surveyed did not emphasise defence as part of their strategy, it, too, is another area for research—in the US, one of the countries studied,

the Department of Defense was responsible for about 85 percent of the total federal funding for AI in 2022.⁷⁷

Space

AI has found extensive application in space operations and exploration, and is already paying dividends in the areas of autonomous navigation and planetary exploration. NASA, for instance, employed the use of an ML navigation system called AutoNav in its Spirit and Opportunity rovers, deployed on Mars in 2004. The AI system helped analyse terrains, plan routes, and avoid obstacles.⁷⁸ The agency's Curiosity rover, which landed in 2012, uses the Autonomous Exploration for Gathering Increased Science (AEGIS) algorithm along with AutoNav to identify rock formations.⁷⁹ AutoNav has since evolved and is now being used in the NASA's Mars 2020 Mission.⁸⁰

The practice of cosmology, which involves processing and analysing massive amounts of data collected from satellites, telescopes, spacecrafts and rovers, is another domain where AI's capabilities in identifying and extracting patterns from large datasets pay huge dividends. The AI in NASA's Kepler telescope, for instance, helped detect new exoplanets such as Kepler-90i and Kepler-1649c.⁸¹ Similarly, NASA's James Webb Space Telescope, launched in 2021, uses an ML model called Morpheus, which aids in detecting and classifying galaxies in deep space.⁸²

AI is also being used during the landing and take-off of spacecrafts to automate engine operations and manage functions such as the deployment of landing gear. SpaceX uses an AI autopilot system that enables autonomous navigation and control, processes real-time sensor data, and uses ML for predictive analytics, to land its Falcon 9 rocket.⁸³ AI ability to model and assess a wide range of mission parameters, making it possible to predict the potential outcomes of different courses of action, means its use case extends to planning space missions.

India has managed to successfully integrate AI into its space programme. The Chandrayaan-2 mission employed the Pragyan rover, which used AI algorithms to trace water and other minerals on the lunar surface and send pictures for research and examination.⁸⁴ AI played a crucial role in the Chandrayaan-3 mission as well, by assisting the mission's Lander Hazard Detection and Avoidance Camera in assessing the lunar topography for obstacles that would otherwise have gotten in the way of a soft landing.⁸⁵

Military

Unlike some important military innovations of the past such as the longbow, gunpowder, or the tank, which had relatively specific uses, AI is a general-purpose technology with an array of applications. More akin to the advent of electricity, which generated advances in lighting, heating, transportation, and communications, AI will diffuse across many other technologies, greatly increasing their capabilities and effectiveness.⁸⁶ There has already been a proliferation in research and development pursuing a variety of military uses for AI, including in autonomous vehicles and weapons systems, intelligence collection, predictive logistics, cybersecurity, and command and control.⁸⁷

The domain of warfare involves processing vast amounts of data and information. Intelligence, surveillance, and reconnaissance assets provide data on enemy forces through a combination of sources, such as imagery, video feeds, signal intercepts, and electromagnetic detections. In addition, friendly forces provide status updates and requests for support over a variety of command-and-control systems. Other factors, such as changes to the weather, the presence of civilians on the battlefield, or the introduction of disinformation, add further complexity to the operational environment.⁸⁸ This is where AI can provide enormous benefits, given its ability to process huge amounts of data at unprecedented speeds.

AI systems can greatly accelerate the military's 'observe, orient, decide, and act' loop by increasing situational awareness, rapidly processing large amounts of information, calculating decision options, and automating operations.⁸⁹ Intelligence analysts can use AI to filter through scores of images and videos to locate enemy forces.

Operators can employ autonomous swarms of drones to overwhelm enemy defences. Logisticians can use data analytics to optimise resupply missions or equipment maintenance. Military planners can use LLMs to draft operations orders and generate decision briefs. Cyber warriors can leverage ML to identify anomalies and deny network intrusions by adversaries.⁹⁰

The Indian military is implementing multiple initiatives to harness AI. The Defence AI Council (DAIC) and the Defence AI Project Agency (DAIPA) were both established in 2019.⁹¹ The DAIC provides the strategic direction towards AI driven transformation in defence. It also offers guidance in addressing issues related to data sharing, enables strategic partnership with industry, decides acquisitions of technology, reviews ethical, safe and privacy assured usage of AI in defence, and sets policies in partnership with government institutions and industries.⁹² The DAIPA's mandate is to evolve and adopt standards for the technology development and delivery process of AI projects. It also reviews the adoption plan of AI-led and AI-enabled systems and processes with user groups.

The Ministry of Defence has provisioned an annual budgetary allocation of INR 1 billion to DAIPA for a period of five years from 2019, for taking up AI projects, setting up AI-related infrastructure, preparing AI-related data and capacity building.⁹³ Each defence service has also earmarked INR 1 billion per year for AI-specific application development in the same period.⁹⁴ The Indian Army has established the Signals Technology Evaluation and Adaptation Group, which is envisioned as an elite unit that will focus on nurturing and developing critical technology domains, including AI and ML.⁹⁵

Finance

AI is now being extensively used in the finance industry to analyse data, automate tasks, and improve decision-making.⁹⁶ AI is being employed in credit scoring and risk assessment since it can analyse vast amounts of data, including social media activity and other online behaviour, to assess a customer's creditworthiness and make more accurate credit decisions.⁹⁷ AI algorithms can also prevent financial crime, such as fraud and cyberattacks, by identifying unusual patterns in

financial transactions. This helps improve security in activities such as online banking and credit card transactions.⁹⁸ AI can also be used to develop trading algorithms that analyse market trends and historical data to make decisions and execute trades.⁹⁹

As of 2022, India possessed the largest number of digital banking users in the world, making the application of AI incredibly beneficial for the financial sector. AI is supporting the improved delivery of banking services and products in the country.¹⁰⁰ The State Bank of India, for instance, uses an AI-based solution to scan cameras installed in its branches that capture the facial expressions of customers to give real-time feedback of their response.¹⁰¹ HDFC bank has deployed an AI-chatbot called Eva, Electronic Virtual Assistant, built by Bengaluru-based Senseforth AI Research.¹⁰² ICICI Bank also offers an AI-chatbot called iPal, in addition to having developed a software robotics platform that leverages AI features such as facial and voice recognition, NLP, ML, and bots.¹⁰³

Agriculture

The agriculture sector is turning to AI technology to cultivate healthier crops, manage pests, monitor soil and growing conditions, analyse data for farmers, and enhance the management of the food supply chain.¹⁰⁴ AI can assist farmers at every stage of crop cultivation. At the time of sowing, it can help farmers choose the optimum seed for a particular weather scenario and the optimal time to plant the seed. Intelligent equipment can also calculate the spacing needed between seeds and the maximum planting depth. Once crops are planted, it offers useful weather forecast data, aids in understanding soil qualities and helps farmers by suggesting the nutrients they should apply to increase the quality of the soil and enhance yield quality and quantity.¹⁰⁵

Of the AI-based technologies to make monitoring of crop and soil health easier, hyperspectral imaging and 3D laser scanning are the leading techniques.¹⁰⁶ Using AI, farmers can thus access advanced data and analytics tools that foster better farming, improve efficiencies,

and reduce waste in biofuel and food production, minimising negative environmental impacts.

India is employing AI to address various challenges in the agricultural sector. For one, the Ministry of Agriculture and Farmers' Welfare has launched the Kisan e-Mitra, an AI chatbot that assists farmers with queries related to the PM Kisan Samman Nidhi Scheme, which is a Central scheme that provides income support to all land-holding farmer families.¹⁰⁷ It supports multiple languages and is being upgraded to assist with other government programmes. Second, the National Pest Surveillance System uses AI and ML to mitigate the loss of produce due to climate change, and enable timely intervention for healthier crops.¹⁰⁸ Third, AI analytics—that process satellite, weather and soil moisture datasets—are being harnessed to assess and monitor the health of crops like rice and wheat.¹⁰⁹ In addition, applications such as the AI Sowing App are being utilised in Karnataka and Andhra Pradesh to provide farmers with information on optimal sowing dates and depths.¹¹⁰

Healthcare and Medicine

AI has made strides in healthcare, and has the potential to transform nearly every aspect of it. ML algorithms are now facilitating the early detection of diseases such as cancer, while also providing more accurate diagnoses.¹¹¹ The integration of AI with wearables and Internet of Things (IoT)-enabled devices is being applied to oversee early-stage heart disease and monitor patient data such as blood pressure and glucose levels, thereby helping healthcare providers manage life-threatening and chronic conditions more effectively.¹¹² AI virtual assistants and chatbots are being used to help answer questions about medications, forward reports to doctors or surgeons, help patients schedule visits with physicians, and improve mental healthcare by engaging users in therapeutic conversations.¹¹³ This also provides clinical staff more time to focus on direct patient care.

In India, AI is playing a pivotal role in transforming the public health landscape. The Government of India's eSanjeevani, a

national telemedicine service, has enhanced access to healthcare, particularly in rural areas, allowing patients to receive consultations and diagnostics without travelling long distances.¹¹⁴ India has developed technologically advanced smart digitalised health and wellness centres (DHWC) that utilise AI capabilities to prioritise preventive health measures, streamline data management and improve patient experiences.¹¹⁵ To enable India's push to Comprehensive Primary Healthcare (CPHC), the Ayushman Bharat Digital Mission (ABDM) has been developed to enable the DHWCs.¹¹⁶ The ABDM platform has been developed with the aim of creating an online platform enabling interoperability of health data within the health ecosystem to create longitudinal electronic health records (EHR) of every citizen. It allows for the tracking of patients using RFID tags and for the monitoring of vitals using IoT wearables. Advanced healthcare plug-ins (such as point of care devices) are integrated into the EHR application to auto-populate health vitals. A pilot implementation of the CPHC platform, the eSwasthya Dham, is being undertaken by the Uttarakhand government.¹¹⁷

Clean Energy

AI has potential to accelerate the global quest for clean energy. AI can process vast amounts of data from satellites, sensors and weather-monitoring stations, which can be used to predict variables like solar radiation and wind speed.¹¹⁸ This allows for the accurate forecasting of renewable energy generation, thereby mitigating the impact of intermittent energy supply. AI can also analyse massive tracts of accumulated consumer data to forecast consumer demand for electricity, which can aid in preventing supply disruptions and blackouts by balancing demand and supply needs.¹¹⁹ The grid management and maintenance of solar panels has improved with the use of AI, since it can offer round-the-clock monitoring of temperatures, irradiance, power output, operational loads and other relevant parameters, while swiftly detecting any anomalies therein.¹²⁰

AI is playing a prominent role in augmenting India's renewable energy sector. Tata Power is using AI to predict the solar energy production from its factory's solar power plant.¹²¹ ReNew

Power uses AI to increase the efficiency of wind turbines. The Power Grid Corporation of India is using AI to improve power grid management.¹²²

Coastal Zone Management

Coastal zone management is an important UN sustainable development goal, and serves as another area where AI can play an important role, particularly in monitoring marine and coastal environments. Belize has taken significant steps in this regard, with the Belize Coastal Zone Management Authority & Institute adopting AI for the Belize National Marine Habitat Map project since late 2020.¹²³ The project uses Microsoft Azure for ML-based mapping of the Essential Biodiversity Variable (EBV) of ecosystem extent and fragmentation. The initiative has adopted the use of ML to help update Belize's 1997 version of its National Marine Habitat Map. The updated data provides revised estimates of the status of Belize's coastal and marine ecosystems and will come into use in formulating a revised National Integrated Coastal Zone Management Plan.¹²⁴ The use of unmanned aerial systems and marine drones for beach litter recognition and underwater litter detection has also been proposed by multiple research groups.¹²⁵

The Future of AI: Small Language Models

What is the direction in which the development of AI is headed and what does its future look like? Novel innovations like the development of small language models (SLMs) may offer an answer. While SLMs offer an exciting new avenue from a developer perspective, they also indicate a way forward for AI development and innovation in the developing nations of the Global South.

Despite the incredible advancements made by generative AI in training LLMs, it is still a time- and resource-intensive task requiring significant

investments and compute power. GPT-3, for instance, requires 175 billion parameters,¹²⁶ Meta's Llama-3.1 405 billion parameters,¹²⁷ while reports suggest GPT-4 needs over a trillion parameters.¹²⁸ LLMs also suffer from a lack of customisability and are not cost-effective for businesses or individuals whose requirements are limited to specific tasks. Furthermore, LLMs require large, high-quality and diverse datasets that are time-consuming and expensive to acquire and preprocess.¹²⁹ The fact that LLMs often employ public datasets also makes them susceptible to bias, in addition to the problem of 'hallucinations,' which leads to inaccurate responses and may make them prone to security risks.

Consequently, training AI models has been a luxury that countries with lesser resources, or smaller businesses and the general masses, cannot afford. This has led to the emergence of small language models (SLMs). These can be used to create smaller, domain-specific or boutique models, tailored to execute customised operations suited for smaller enterprises or, in principle, even individuals. While no universal definition of SLMs exists currently, they are usually five to ten times smaller than LLMs in terms of the number of parameters they need.¹³⁰ As a result, they require lesser computational power and memory, and also consume much lesser energy. The lower resource requirement has the added benefit of democratising AI, allowing smaller corporations, teams and individual researchers to train AI models.¹³¹ For instance, Meta's Llama 2 7B consists of seven billion parameters as compared to its larger counterpart Llama 2, which contains 34 billion parameters.¹³² Its Alpaca 7B is a fine-tuned version of the Llama 2 7B and required less than US\$600 to build.¹³³ Other examples of SLMs include the Stable Beluga 7B, X Gen, MPT, Falcon 7B, and Zephyr, all of which are built on seven billion parameters, while Microsoft's Phi-2 is based on 13 billion parameters.¹³⁴ SLMs with parameter sizes below one billion are also in existence, DistillBERT, TinyBERT and T5-Small,

serving as some prominent examples.¹³⁵ However, their utility is quite limited at the moment.

SLMs are ideal for IoT edge devices and mobile phones. While users can access an LLM on the cloud using mobile devices, it still requires a high-speed internet connection, with the performance being limited by the speed of the connection.¹³⁶ In developing countries like India that lack universal broadband connectivity, this issue becomes particularly important. Individuals or small organisations without access to a high-speed internet connection have no hope of exploiting most of the benefits of an LLM. On the other hand, an SLM that is small enough to fit on a mobile phone but still efficient and powerful enough to perform tasks accurately and quickly, offers a viable solution to this problem, and would provide increased access to AI models for the masses, particularly in rural areas. Restricting computation on an individual device also saves costs by not sending data to be processed in the cloud.¹³⁷ Furthermore, the model can be grounded and trained on the data on an individual's phone and can be personalised to their needs. Keeping customer data secured within its own secured platform is also desirable from a security and privacy perspective, particularly for sensitive use cases like banking and healthcare.¹³⁸

One drawback of current SLMs is that despite being relatively smaller than LLMs, they still require sizeable datasets. This problem can be overcome using a few clever techniques. One of these is knowledge distillation, which involves transferring knowledge from a pre-trained LLM to a smaller model, capturing its core capabilities without the full complexity.¹³⁹ The removal of the unnecessary parts of a larger model along with a reduction in the precision of its weights can also be employed to create a domain-specific SLM.¹⁴⁰ Building models that allow for easy swapping of SLMs is another technique that can be employed to address the problem of large datasets, while also handling any unexpected changes to the platform.¹⁴¹

The Future of SLMs

How small can SLMs be in principle? Many investigations have found that modern training methods can impart basic language competencies in models with no more than 10 million parameters.¹⁴² For example, an eight-million parameter model

released in 2023 attained 59 percent accuracy on the established natural language understanding benchmark GLUE.¹⁴³

Performance improves as an AI model's capacity grows. A 2023 study found that useful capability thresholds for different tasks—from reasoning to translation—were consistently passed once language models hit about 60 million parameters.¹⁴⁴ However, returns diminished after the 200–300 million mark; any additional capacity only led to incremental performance gains.¹⁴⁵

These findings suggest even mid-sized language models can hit reasonable competence across many language processing applications, provided they are exposed to enough of the right training data. Performance then reaches a plateau where cramming further data into the model does not provide much additional value. The sweet spot for commercially deployable SLMs likely rests around this plateau zone, balancing broad-spectrum ability with concise efficiency. Specialised SLMs tuned deeply rather than broadly may require even lesser capacity to excel at niche tasks.¹⁴⁶

Consequently, the objective now is to bring SLMs down to a few million parameters, which in principle, is completely within the realm of possibility, particularly for domain-specific tasks. This will make it possible for users to train AI


models in resource-constrained environments like mobile phones, without the need for any internet connectivity. This will go a long way in democratising AI models for the general public, particularly in developing nations like India.

Conclusion: AI Development in the Global South and the Role of Regulation

With the massive resources and investment required for training large-scale AI models, the technology has remained confined to the more advanced economies of the world, while the relatively weaker economies of the Global South have lagged behind. Consequently, the Global South is seen merely as a market for products developed elsewhere, which may not necessarily address its needs. However, as AI innovations continue to emerge from countries like India, this perception may no longer be tenable. SLMs may emerge as an added boon in this regard since they would enable democratisation of AI models.

Excitement over the future of AI, however, must be tempered with an attention to mitigating its risks. While the advancement in AI over the preceding decade has been nothing short of extraordinary, it has also brought to light lingering issues with potentially dangerous consequences. These include, but are not limited to, the AI black box problem, hallucinations, and biased datasets. Such issues foreground the importance of AI regulation and the need to develop legal frameworks that encourage the responsible and ethical use of AI while embedding these principles into the development of AI itself.

The pursuit of AI regulation has led to the emergence of two broadly distinct priorities: that of innovation, being actively encouraged by the US, which is home to Big Tech corporations like Microsoft and OpenAI, and regulation, being propagated by the EU by virtue of its AI Act, passed in 2024. In this backdrop, Global South nations are faced with the dilemma of deciding which of these paths is most appropriate for them. Though regulation may seem appealing from an ethical standpoint, overregulation can stifle innovation while making domestic developers less competitive in comparison to their Western counterparts.

On the other hand, letting innovation reign free without the requisite guards poses ethical risks and would be highly irresponsible. Consequently, a middle ground between the two would be the most appropriate for developing nations like India. Such a path would entail addressing primary concerns like mitigating bias in AI systems, while also giving developers room to innovate. This would help in establishing Global South countries as a hub for AI development and push them towards self-sufficiency. 

Endnotes

- 1 “AI Versus Machine Learning Versus Deep Learning Versus Neural Networks: What’s The Difference?,” *IBM*, July 6, 2023, <https://www.ibm.com/think/topics/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks>.
- 2 ” “AI Versus Machine Learning Versus Deep Learning Versus Neural Networks: What’s The Difference?”
- 3 “AI Versus Machine Learning Versus Deep Learning Versus Neural Networks: What’s The Difference?”
- 4 “AI Versus Machine Learning Versus Deep Learning Versus Neural Networks: What’s The Difference?”
- 5 “What Is a Neural Network?,” *IBM*, <https://www.ibm.com/topics/neural-networks>.
- 6 Larry Hardesty, “Explained: Neural Networks,” *MIT News*, April 14, 2017, <https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414>.
- 7 Hardesty, “Explained: Neural Networks”
- 8 Hardesty, “Explained: Neural Networks”
- 9 “What Is a Neural Network?”
- 10 “What Is Deep Learning?,” *IBM*, <https://www.ibm.com/topics/deep-learning>.
- 11 “AI Versus Machine Learning Versus Deep Learning Versus Neural Networks: What’s The Difference?”
- 12 “What Is Deep Learning?”
- 13 “What Is Deep Learning?”
- 14 “What Is Deep Learning?”
- 15 “What Is a Neural Network?”
- 16 [qure.ai](https://www.qure.ai/), <https://www.qure.ai/>
- 17 [Whiterabbit.ai](https://www.whiterabbit.ai/), <https://www.whiterabbit.ai/>
- 18 [Nanonets](https://nanonets.com/), <https://nanonets.com/>
- 19 [ParallelDots](https://www.paralleldots.com/), <https://www.paralleldots.com/>
- 20 [Tracxn](https://tracxn.com/d/companies), <https://tracxn.com/d/companies>
- 21 “What Is Machine Learning?,” *IBM*, <https://www.ibm.com/topics/machine-learning>.
- 22 “What Is Machine Learning?”

- 23 “Deep Learning vs Machine Learning: The Ultimate Battle,” *Turing*, <https://www.turing.com/kb/ultimate-battle-between-deep-learning-and-machine-learning>.
- 24 “What Is Machine Learning?”
- 25 “What Is Machine Learning?”
- 26 uniphore, <https://www.uniphore.com/>
- 27 Eightfold, <https://eightfold.ai/>
- 28 SigTuple, <https://www.sigtuple.com/>
- 29 haptik, <https://www.haptik.ai/>
- 30 Tracxn, <https://tracxn.com/d/companies>
- 31 “What Is Natural Language Processing (NLP)?” *Amazon Web Services*, <https://aws.amazon.com/what-is/nlp/>.
- 32 Manish Goyal, Shobhit Varshney and Eniko Rozsa, “What Is Generative AI, What Are Foundation Models, And Why Do They Matter?,” *IBM*, March 8, 2023, <https://www.ibm.com/blog/what-is-generative-ai-what-are-foundation-models-and-why-do-they-matter/>.
- 33 Mike Murphy, “What Are Foundation Models?,” *IBM*, May 9, 2022, <https://research.ibm.com/blog/what-are-foundation-models>.
- 34 Murphy, “What Are Foundation Models?”
- 35 Murphy, “What Are Foundation Models?”
- 36 Goyal et al., “What Is Generative AI, What Are Foundation Models, And Why Do They Matter?”
- 37 Goyal et al., “What Is Generative AI, What Are Foundation Models, And Why Do They Matter?”
- 38 Hardesty, “Explained: Neural Networks”
- 39 Himanshu Singh, “Everything You Need To Know About Hardware Requirements For Machine Learning,” *eInfochips*, April 24, 2024, <https://www.einfochips.com/blog/everything-you-need-to-know-about-hardware-requirements-for-machine-learning/>.
- 40 Singh, “Everything You Need To Know About Hardware Requirements For Machine Learning”
- 41 Ben Wodecki, “Meta Reveals GPU Clusters Used To Train Llama 3,” *AI Business*, March 12, 2024, <https://aibusiness.com/verticals/meta-reveals-gpu-clusters-used-to-train-llama-3>
- 42 Wodecki, “Meta Reveals GPU Clusters Used To Train Llama 3”
- 43 Andrew Froehlich, “Building Networks For AI Workloads,” *TechTarget*, April 11, 2024, <https://www.techtarget.com/searchnetworking/tip/Building-networks-for-AI-workloads>.
- 44 Froehlich, “Building Networks For AI Workloads”

- 45 Suhas Nayak, "Scaling AI Infrastructure With High-Speed Optical Connectivity," *Marvell*, June 27, 2023, <https://www.marvell.com/blogs/scaling-ai-infrastructure-with-high-speed-optical-connectivity.html>.
- 46 "Infrastructure: Machine Learning Hardware Requirements," *c3.ai*, <https://c3.ai/introduction-what-is-machine-learning/machine-learning-hardware-requirements/>.
- 47 Singh, "Everything You Need To Know About Hardware Requirements For Machine Learning"
- 48 "Infrastructure: Machine Learning Hardware Requirements"
- 49 "Infrastructure: Machine Learning Hardware Requirements"
- 50 Lauren Leffer, "The AI Boom Could Use a Shocking Amount Of Electricity," *Scientific American*, October 13, 2023, <https://www.scientificamerican.com/article/the-ai-boom-could-use-a-shocking-amount-of-electricity/>.
- 51 Leffer, "The AI Boom Could Use a Shocking Amount Of Electricity"
- 52 Leffer, "The AI Boom Could Use a Shocking Amount Of Electricity"
- 53 Singh, "Everything You Need To Know About Hardware Requirements For Machine Learning"
- 54 Aamir Aftab, "AI Chips and Hardware Acceleration: ASICs, TPUs, GPUs," *Medium*, January 4, 2024, <https://medium.com/@aamiraftabcloud/ai-chips-and-hardware-acceleration-asics-tpus-gpus-a881993bf92f>.
- 55 Aftab, "AI Chips and Hardware Acceleration: ASICs, TPUs, GPUs"
- 56 "FPGA vs. GPU For Deep Learning," *Intel*, <https://www.intel.com/content/www/us/en/artificial-intelligence/programmable/fpga-gpu.html>.
- 57 "FPGA vs. GPU For Deep Learning"
- 58 Saif M. Khan, "AI Chips: What They Are And Why They Matter," *Center for Security and Emerging Technology*, April, 2020, <https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/>.
- 59 Kevin Krewell, "IBM's Cloud AI Supercomputer Vela Builds AI Foundation Models For Enterprise," *Forbes*, February 23, 2023, <https://www.forbes.com/sites/tiriasresearch/2023/02/23/ibms-cloud-ai-supercomputer-vela-builds-ai-foundation-models-for-enterprise/?sh=7162ad752f4a>.
- 60 Krewell, "IBM's Cloud AI Supercomputer Vela Builds AI Foundation Models For Enterprise"
- 61 Krewell, "IBM's Cloud AI Supercomputer Vela Builds AI Foundation Models For Enterprise"
- 62 Krewell, "IBM's Cloud AI Supercomputer Vela Builds AI Foundation Models For Enterprise"
- 63 Arthur Cole, "How Will AI Change The User Interface?" *Techopedia*, October 5, 2023, <https://www.techopedia.com/how-will-ai-change-the-user-interface>.
- 64 Jakob Nielsen, "AI: First New UI Paradigm In 60 Years," *Nielsen Norman Group*, June 18, 2023, <https://www.nngroup.com/articles/ai-paradigm/>.

- 65 Maximillian Piras, “When Words Cannot Describe: Designing For AI Beyond Conversational Interfaces,” *Smashing Magazine*, February 2, 2024, <https://www.smashingmagazine.com/2024/02/designing-ai-beyond-conversational-interfaces/>.
- 66 Cole, “How Will AI Change The User Interface?”
- 67 Cole, “How Will AI Change The User Interface?”
- 68 Piras, “When Words Cannot Describe: Designing For AI Beyond Conversational Interfaces”
- 69 Piras, “When Words Cannot Describe: Designing For AI Beyond Conversational Interfaces”
- 70 Nielsen, “AI: First New UI Paradigm In 60 Years”
- 71 Nielsen, “AI: First New UI Paradigm In 60 Years”
- 72 Nielsen, “AI: First New UI Paradigm In 60 Years”
- 73 Cole, “How Will AI Change The User Interface?”
- 74 Bernard Marr, “AI’s Next Frontier: Are Brain-Computer Interfaces the Future Of Communication?,” *Forbes*, August 11, 2023, <https://www.forbes.com/sites/bernardmarr/2023/08/11/ais-next-frontier-are-brain-computer-interfaces-the-future-of-communication/?sh=5fbab8c851d9>.
- 75 James S. Denford, Gregory S. Dawson and Kevin C. Desouza, “A Cluster Analysis Of National AI Strategies,” *Brookings*, December 13, 2023, <https://www.brookings.edu/articles/a-cluster-analysis-of-national-ai-strategies/>.
- 76 “National Strategy For Artificial Intelligence,” *NITI Aayog*, June, 2018, <https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf>.
- 77 Denford et al., “A Cluster Analysis Of National AI Strategies”
- 78 “NASA’s Perseverance Rover Hightails It To Martian Delta,” *Jet Propulsion Laboratory*, March 18, 2022, <https://www.jpl.nasa.gov/news/nasas-perseverance-rover-hightails-it-to-martian-delta/>.
- 79 “NASA’s Perseverance Rover Hightails It To Martian Delta”
- 80 “AutoNav Drives Perseverance Forward,” *NASA*, April 19, 2022, <https://science.nasa.gov/resource/autonav-drives-perseverance-forward/>.
- 81 Bernard Marr, “Artificial Intelligence In Space: The Amazing Ways Machine Learning Is Helping To Unravel the Mysteries Of the Universe,” *Forbes*, April 10, 2023, <https://www.forbes.com/sites/bernardmarr/2023/04/10/artificial-intelligence-in-space-the-amazing-ways-machine-learning-is-helping-to-unravel-the-mysteries-of-the-universe/>.
- 82 Ben Wodecki, “AI To Help NASA’s James Webb Telescope Map the Stars,” *AI Business*, July 12, 2022, <https://aibusiness.com/verticals/ai-to-help-nasa-s-james-webb-telescope-map-the-stars>.
- 83 Marr, “Artificial Intelligence In Space: The Amazing Ways Machine Learning Is Helping To Unravel the Mysteries Of the Universe”
- 84 Leslie D’Monte, “Chandrayaan-2 Pragyan Shows How AI Is Helping Space Exploration,” *Mint*, September 6, 2019, <https://www.livemint.com/technology/tech-news/chandrayaan-2-pragyan-shows-how-ai-is-helping-space-exploration-1567764065716.html>.

- 85 Milin Stanly, "Chandrayaan 3: How AI Drove a Historic Landing On the Moon," *INDIAai*, September 6, 2024, <https://indiaai.gov.in/article/chandrayaan-3-how-ai-drove-a-historic-landing-on-the-moon>.
- 86 Col. Joshua Glonek, "The Coming Military AI Revolution," *Military Review*, May-June 2024, <https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/May-June-2024/MJ-24-Glonek/>
- 87 Glonek, "The Coming Military AI Revolution"
- 88 Glonek, "The Coming Military AI Revolution"
- 89 Glonek, "The Coming Military AI Revolution"
- 90 Glonek, "The Coming Military AI Revolution"
- 91 "Enhancement Of Capabilities Of AI Technology," Ministry of Defence, Government of India, August 1, 2022, <https://www.pib.gov.in/PressReleasePage.aspx?PRID=1846937>.
- 92 Lt. Gen. Deependra Singh Hooda, "Implementing Artificial Intelligence In the Indian Military," *Delhi Policy Group*, February 16, 2023, <https://www.delhipolicygroup.org/publication/policy-briefs/implementing-artificial-intelligence-in-the-indian-military.html>.
- 93 Hooda, "Implementing Artificial Intelligence In the Indian Military"
- 94 Hooda, "Implementing Artificial Intelligence In the Indian Military"
- 95 Kartik Bommakanti, "The STEAG: A New Development In the Army's Technological Capabilities," *Observer Research Foundation*, April 30, 2024, <https://www.orfonline.org/expert-speak/the-steag-a-new-development-in-the-armys-technological-capabilities>.
- 96 Matthew Finio and Amanda Downie, "What Is AI In Finance?," *IBM*, December 8, 2023, <https://www.ibm.com/topics/artificial-intelligence-finance>.
- 97 Finio and Downie, "What Is AI In Finance?"
- 98 Finio and Downie, "What Is AI In Finance?"
- 99 Bernard Marr, "15 Amazing Real-World Applications Of AI Everyone Should Know About," *Forbes*, May 10, 2023, <https://www.forbes.com/sites/bernardmarr/2023/05/10/15-amazing-real-world-applications-of-ai-everyone-should-know-about/?sh=5862170785e8>.
- 100 Amitabh Chaudhry, "Here Are 4 Ways AI Is Streamlining Banking In India," *World Economic Forum*, December 20, 2023, <https://www.weforum.org/agenda/2023/12/how-ai-can-streamline-indian-banking/>.
- 101 Ayushman Baruah, "AI Applications In the Top 4 Indian Banks," *emerj*, February 27, 2020, <https://emerj.com/ai-sector-overviews/ai-applications-in-the-top-4-indian-banks/>.
- 102 Baruah, "AI Applications In the Top 4 Indian Banks"
- 103 Baruah, "AI Applications In the Top 4 Indian Banks"

- 104 Mohammad Javed, Abid Haleem, Ibrahim Haleem Khan and Rajiv Suman, “Understanding the Potential Applications Of Artificial Intelligence in Agriculture Sector,” *Advanced Agrochem*, Volume 2, Issue 1, Pages 15-30, March, 2023, <https://www.sciencedirect.com/science/article/pii/S277323712200020X>.
- 105 Javed et al., “Understanding the Potential Applications Of Artificial Intelligence in Agriculture Sector”
- 106 Javed et al., “Understanding the Potential Applications Of Artificial Intelligence in Agriculture Sector”
- 107 “Use Of AI To Tackle Problems In Agriculture,” Ministry of Agriculture & Farmers Welfare, Government of India, February 2, 2024, <https://pib.gov.in/PressReleaseIframePage.aspx?PRID=2002010>.
- 108 “Use Of AI To Tackle Problems In Agriculture”
- 109 “Use Of AI To Tackle Problems In Agriculture”
- 110 “India Is At the Cusp Of a Farming Revolution Through AI,” *DownToEarth*, January 23, 2024, <https://www.downtoearth.org.in/agriculture/india-is-at-the-cusp-of-a-farming-revolution-through-ai-94055>.
- 111 “No Longer Science Fiction, Ai And Robotics Are Transforming Healthcare,” *PwC*, <https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health/transforming-healthcare.html>.
- 112 Marr, “15 Amazing Real-World Applications Of AI Everyone Should Know About”
- 113 “The Benefits Of AI In Healthcare,” *IBM*, July 11, 2023, <https://www.ibm.com/think/insights/ai-healthcare-benefits>.
- 114 Dr Rakesh Kumar, “How Digital Tech And AI Are Revolutionising Primary Health Care In India,” *Business Standard*, July 11, 2024, https://www.business-standard.com/health/how-digital-tech-and-ai-are-revolutionising-primary-health-care-in-india-124071100212_1.html.
- 115 Kumar, “How Digital Tech And AI Are Revolutionising Primary Health Care In India”
- 116 Kumar, “How Digital Tech And AI Are Revolutionising Primary Health Care In India”
- 117 Kumar, “How Digital Tech And AI Are Revolutionising Primary Health Care In India”
- 118 Sumant Sinha, “AI Can Power the Green Energy Transition,” *Forbes*, February 26, 2024, <https://www.forbes.com/sites/sumantsinha/2024/02/26/ai-can-power-the-green-energy-transition/>.
- 119 Sinha, “AI Can Power the Green Energy Transition”
- 120 Luiz Avelar and Guy Borthwick, “Sun, Sensors and Silicon: How AI Is Revolutionizing Solar Farms,” *World Economic Forum*, August 2, 2024, <https://www.weforum.org/agenda/2024/08/how-ai-can-help-revolutionize-solar-power/>.
- 121 Ankur Kumar, “Assessing the Impact Of Artificial Intelligence On Renewable Energy In India,” *Earth5R*, <https://earth5r.org/assessing-the-impact-of-artificial-intelligence-on-renewable-energy-in-india/>.
- 122 Kumar, “Assessing the Impact Of Artificial Intelligence On Renewable Energy In India”
- 123 “GEO BON – Microsoft: EBV’s On the Cloud AI For the National Belize Marine Habitat Map,” *The Coastal Zone Management Authority and Institute*, <https://www.coastalzonebelize.org/portfolio/ai-for-belize-marine-habitat-map/>.
- 124 “GEO BON – Microsoft: EBV’s On the Cloud AI For the National Belize Marine Habitat Map”

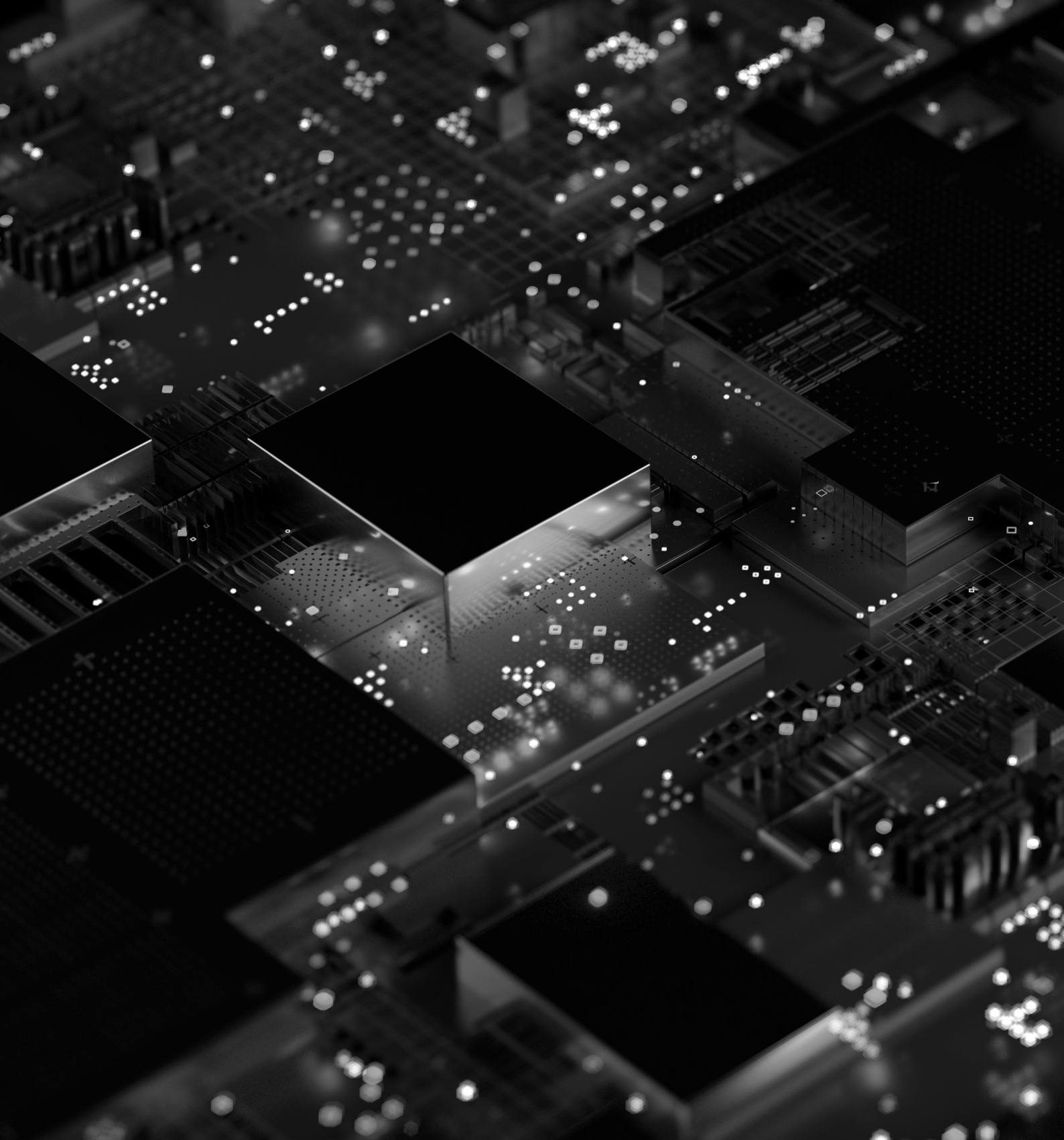
- 125 Gabriela Escobar-Sánchez, Greta Markfort, Mareike Berghald, Lukas Ritzenhofen and Gerald Schernewski, “Aerial and underwater drones for marine litter monitoring in shallow coastal waters: factors influencing item detection and cost-efficiency,” *Environ Monit Assess* 194, 863 (2022), <https://doi.org/10.1007/s10661-022-10519-5>.
- 126 Bijit Ghosh, “The Rise Of Small Language Models – Efficient & Customizable,” *Medium*, November 26, 2023, <https://medium.com/@bijit211987/the-rise-of-small-language-models-efficient-customizable-cb48ddee2aad>.
- 127 Armand Ruiz, Maryam Ashoori and Dave Bergmann, “Meta Releases New Llama 3.1 Models, Including Highly Anticipated 405B Parameter Variant,” *IBM*, July 23, 2024, <https://www.ibm.com/blog/meta-releases-llama-3-1-models-405b-parameter-variant/>.
- 128 Josh Howarth, “Number Of Parameters In GPT-4,” *Exploding Topics*, August 6, 2024, <https://explodingtopics.com/blog/gpt-parameters>.
- 129 Lisa Lee, “Tiny Titans: How Small Language Models Outperform LLMs For Less,” *Salesforce*, June 3, 2024, <https://www.salesforce.com/blog/small-language-models/>
- 130 Tom Taulli, “Small Language Models Gaining Ground At Enterprises,” *AI Business*, January 24, 2024, <https://aibusiness.com/nlp/small-language-models-gaining-ground-at-enterprises>.
- 131 Nagesh Mashette, “Small Language Models (SLMs),” *Medium*, December 12, 2023, <https://medium.com/@nageshmashette32/small-language-models-slms-305597c9edf2>.
- 132 Sandhra Jayan, “9 Best Small Language Models in 2024,” *Analytics India Magazine*, December 7, 2023, <https://analyticsindiamag.com/developers-corner/best-small-language-models/>.
- 133 Jayan, “9 Best Small Language Models in 2024”
- 134 Jayan, “9 Best Small Language Models in 2024”
- 135 Mashette, “Small Language Models (SLMs)”
- 136 Lee, “Tiny Titans: How Small Language Models Outperform LLMs For Less”
- 137 Lee, “Tiny Titans: How Small Language Models Outperform LLMs For Less”
- 138 Lee, “Tiny Titans: How Small Language Models Outperform LLMs For Less”
- 139 Mashette, “Small Language Models (SLMs)”
- 140 Mashette, “Small Language Models (SLMs)”
- 141 Taulli, “Small Language Models Gaining Ground At Enterprises”
- 142 Ghosh, “The Rise Of Small Language Models – Efficient & Customizable”
- 143 Ghosh, “The Rise Of Small Language Models – Efficient & Customizable”
- 144 Ghosh, “The Rise Of Small Language Models – Efficient & Customizable”
- 145 Ghosh, “The Rise Of Small Language Models – Efficient & Customizable”
- 146 Ghosh, “The Rise Of Small Language Models – Efficient & Customizable”

About the Author

***Prateek Tripathi** is Junior Fellow, Centre for Security, Strategy and Technology, ORF.*

Cover photo: Getty Images/Andriy Onufriyenko

Back cover image: Getty Images/Andriy Onufriyenko



Ideas . Forums . Leadership . Impact

**20, Rouse Avenue Institutional Area,
New Delhi - 110 002, INDIA
Ph. : +91-11-35332000. Fax : +91-11-35332005
E-mail: contactus@orfonline.org
Website: www.orfonline.org**