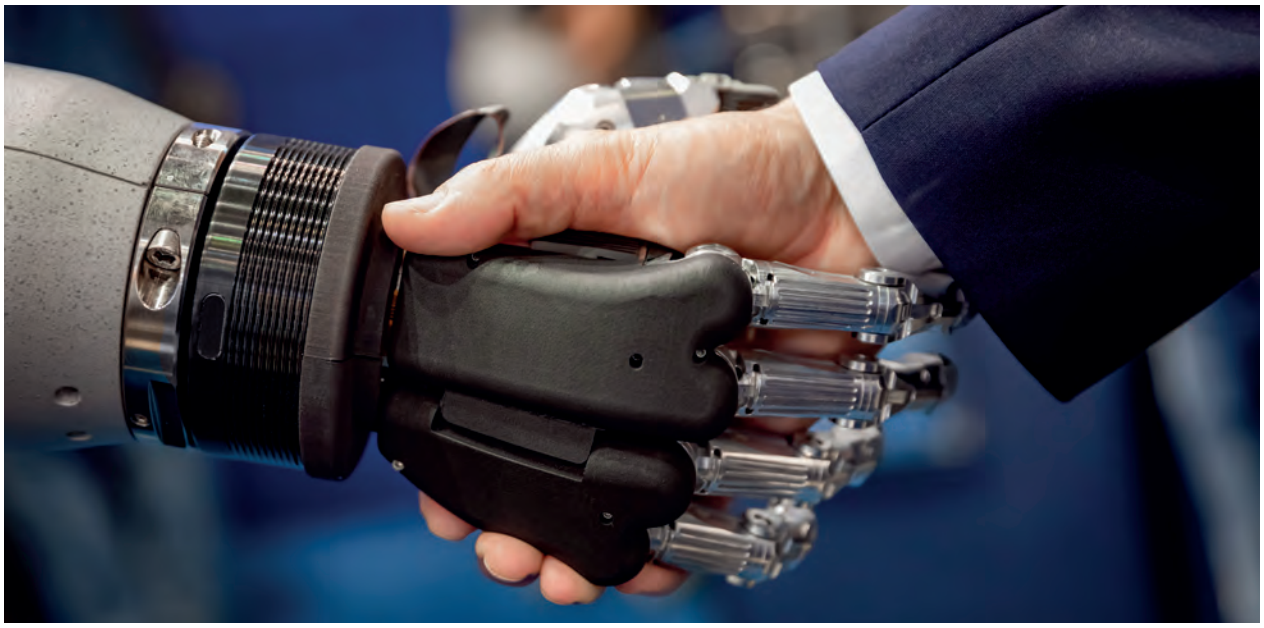
Infosys®
Navigate your next

PERVASIVE INTELLIGENCE FOR A LIVE ENTERPRISE

Infosys® | Knowledge Institute

Contents

Artificial intelligence (AI) has become pervasive in this era of technological advancements. Enterprises are leveraging AI at varying degrees, triggered by pandemic-induced disruptions. AI has evolved from augmented intelligence using classical algorithms to responsible and explainable AI systems using advanced deep learning-based models. Businesses should move across three horizons to evolve as AI-first live enterprises.



Business networks have become unpredictable during global disruptions, specifically COVID-19. The rising globalization of supply chain networks has elevated disruption risks, led by demand volatility and lack of operational control. Enterprises are also cautious of intensifying competition, with faster product launches and growing customer expectations.

The key here is to respond to these changes intelligently—enterprises should be AI-ready to promptly sense the changing dynamics of customers, businesses, partners, and employees. Enterprises need to adopt AI-first strategy to build future-competent systems by creating competitive advantages with better products, services, and business models. Every business process, application, and customer interaction should be AI-efficient to provide a perceptive and even sentient experience.

Shift toward pervasive AI

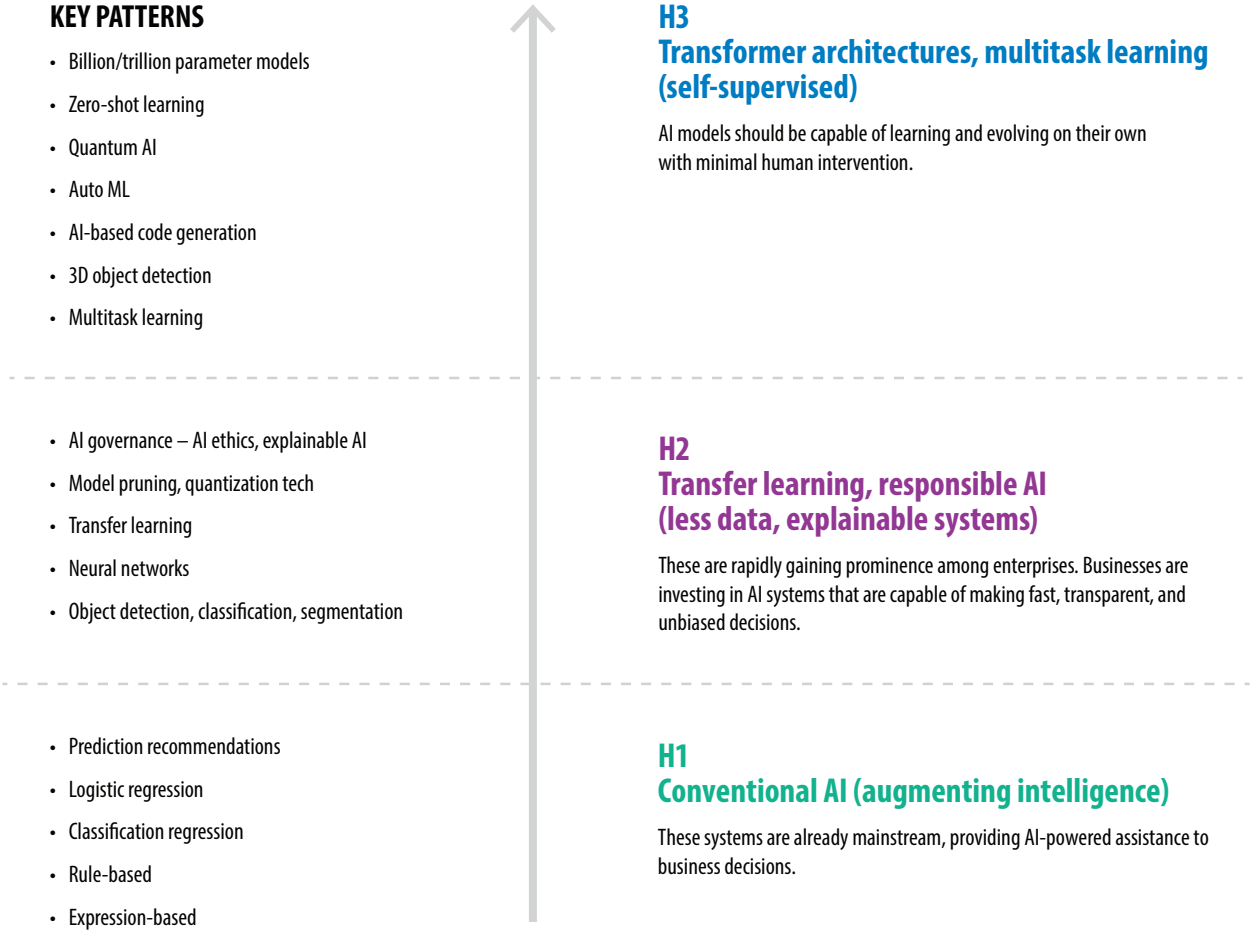
Enterprises should move across three horizons for AI transformation:

Horizon 1 (H1) systems involve augmenting fragmented intelligence into existing systems with customer recommendations and fault identification, usually through classical AI algorithms such as Naive Bayes, support vector machines (SVMs), and random forest.

H2 systems are complex, requiring higher-order generalization, accuracy, and learning capabilities; for example, neural machine translations and conversational insights through deep-learning algorithms.

H3 systems drive organizations toward semi-supervised to unsupervised, transparent, multitask learning systems. Semi-supervised machine intelligence-based examples involve generative capability for textual, audio, and video content; and video-based insights generation, such as activity recognition or summarization. Sensors, cameras, and connected devices deliver rich intelligence at the edge with distributed learning. All this is achieved through well-managed, governed AI systems that are interpretable and explainable at all lifecycle stages. In this paradigm, the role of data engineers is elevated, with a focus on composing automated, modular, and fit-to-address AI approaches. Here, the ability to manage incoming data while monitoring and fixing live applications becomes critical.

Figure 1. Adapt to market dynamics: the three horizons



Source: Infosys

Migrating from H1 to H3 will take enterprises on a transformation journey across the following subdomains:

- AI algorithms
- Natural language processing (NLP)
- Speech
- Computer vision
- AI on edge
- Data engineering
- Responsible AI
- AI platforms

The below key trends under each subdomain can help enterprises transform from fragmented, ad hoc intelligent entities to creative, efficient, and responsibly intelligent organizations.

Figure 2. Key trends across AI subdomains

Source: Infosys

AI ALGORITHMS

Trend 1: Deep neural network architectures help improve generalization and accuracy[1]

Deep-learning algorithms promise higher accuracy and better generalization characteristics than classical algorithms such as SVM, Naive Bayes, and random forest. Enterprise-class problems can be aptly resolved through graphics processing unit (GPU) computing; accessibility of large, labeled data; and fast-paced innovations in deep-learning algorithms. However, the need for a large set of labeled data and the cost of GPU computing are key challenges. Nevertheless, transfer learning-based models, which involve storing knowledge on one domain and then applying to other related problems, have made massive headway in overcoming the limitations of insufficient labeled data and GPU. Also, evolving architecture, such as transformers, has eased certain complex problems in computer vision, NLP, and speech domains.

A large technology company wanted to advance its existing system that worked on certain preconfigured historical rules and policies to moderate user-uploaded content. The company, in partnership with Infosys, developed an AI model for supervised transfer of learning-based deep neural net architecture for vision and text. This helped the company to identify, classify, and isolate any toxic content arriving from user-uploaded forms.

Trend 2: Transition from system 1 to system 2 deep learning[2]

The current state of deep learning-based AI is referred as system 1 deep learning. For example, a person can easily drive in a known vicinity without consciously focusing on directions. However, the same person in an unknown vicinity would require logical reasoning and connections to drive to the destination. These types of problems, which require a combination of reasoning and a sense of "on-the-fly decision-making", are system 2 deep learning.

System 1 deep learning has certain limitations related to generalization capabilities, where these algorithms -

- are not able to accurately work on (detect) unseen data patterns;
- need to have balanced distribution of data in training and testing sets;
- lack continuous learning based on changes in environments in real time, similar to active agents;
- lack logical and reasoning capability to combine high-level semantic concepts, and
- are unable to deal with out-of-distribution (noise) data.

System 2 deep learning resolves some of these challenges by leveraging attention-based architectures and models (the general task of dealing with events over time) and multitask learning (multiple tasks solved at the same time), and incorporating principles of consciousness and meta learning, with an emphasis on unsupervised, zero-shot learning techniques. In zero-shot learning techniques, observed classes in the data are associated with non-observed classes through some form of auxiliary information. This speeds up processing times and increases the efficiency of tasks such as object detection and NLP.

NATURAL LANGUAGE PROCESSING

Enterprises deal with a large database of unstructured documents, images, emails, blogs, voice conversations, and videos, representing a huge opportunity for mining intelligence. NLP is undergoing significant innovations to improve machines' ability to understand (NLU), generate (NLG), and process and derive insights (NLP). Similar to vision and speech, transfer of learning-based text training plays a significant role in getting NLP-based models running quickly.

Transfer of learning-based text training is vital to speed up NLP-based models

The early use of NLP (H1) primarily focused on extracting and representing information as just a bag of words, with hot encoding-based sparse vectors emphasizing on individual words rather than sentence construction, leading to loss of meaning. Also, it extracted named entities (organization, person, location, date, and time) from the text, with nearly no ability to detect/extract custom entities such as currency symbols. In the second phase (H2), deep learning-based word vector models such as GloVe and Word2Vec helped establish word similarity and synonyms without needing to train for every word in the dictionary. The system could also identify custom-named entities such as currency symbols. However, this system was incompetent to deal with out-of-vocabulary words, misspelled words, sentences, paragraphs, and language context-related challenges such as entity coreference resolutions.

The lack of contextual information in embeddings and the reliance on distances between vectors resulted in the evolution of word distance-based biases. For instance, "doctor" and "man" would be more tightly woven together in these models than "doctor" and "woman," given the tendency for these words to be closer together in many research reports and other information sources.

Long-term short-term memory architectures (LSTMs) overcame these limitations by capturing long-range dependency through memory gates (cells) and dealing with coreference resolution problems. However, the lack of parallel processing capability presented the challenge of a long processing time.

Then, attention networks came to the forefront with their capability to focus on a specific part of an input sentence, reducing the training time with parallel processing. Transformers with encoding and decoding architecture are special types of attention networks.[3] They can be trained in parallel, deal with longer

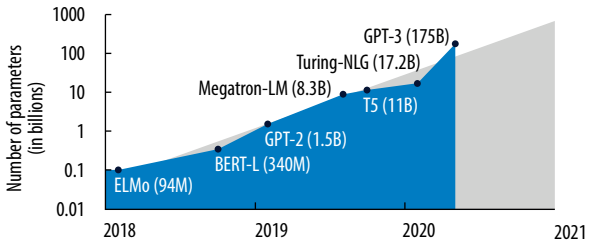
sentences, and are a lot faster to train and predict. They are the architectural building blocks behind all the modern state-of-the-art results provided by various models — Google's BERT,[4] XLNet, and T5; OpenAI's GPT-2;[5] Facebook's RoBERTa; and Baidu's ERNIE 2.0.

Several variations of BERT-based architectures have come up, specifically VideoBERT (a joint model for video and language representation learning) and ViLBERT (a joint model for learning language and vision).

ERNIE 2.0 supports multitask learning and has better accuracy than BERT and XLNet on sixteen standard language tasks and certain other Chinese tasks. SuperGLUE, the advanced version of GLUE with more complex tasks, is the new platform for NLP tasks.[6] Presently, Google's text-to-text transfer transformer (T5) is the leader, with the highest accuracy for all 10 tasks.[7]

OpenAI's generative pre-trained transformer (GPT-3) is by far the latest and the largest model, with 175 billion parameters, excelling in numerous language tasks with few-shot learning (FSL).[8] While GPT-3 is available as a commercial web-API, EleutherAI open sourced a 6 billion parameter model, GPT-J.[9]

Figure 3. The evolution of state-of-the-art NLP models[10]



Source: NVIDIA

H3-based implementations in NLP are shaping up with the usage of previously discussed state-of-the-art transformer-based architectures and models, leveraging contextual and cross-lingual word embeddings and FSL to solve various language tasks.

NLP finds uses in several areas—autoclassification, clustering of documents, extracting key entities and paragraphs, or doing sentiment analysis of text. GPT-3 and GPT-J models are primarily text-in and text-out kinds of models. Therefore, they are also used for source code-related tasks. The next key trend is adopting AI to augment generation, translation, validation, and documentation of code in enterprises. GitHub's CoPilot specializes in code completion[11] and OpenAI's Codex in translating natural languages to code.[12]

Trend 3: Active learning for content intelligence from documents

Enterprises embed information in various types of documents, digital or handwritten, comprising research study documents, know-your-customer (KYC) forms, payslips, and invoices. Here, extracting and systematically digitizing this information is a huge challenge. One advanced technique to derive content intelligence from documents is active learning. An AI classifier examines unlabeled data and picks parts of this data for further human labeling. This active process increases data quality, as the classifier controls data selection and picks only areas that are not optimized for ML. In one such legal use case (labeling contractual clauses), active learning increased data accuracy from 66% to 80%, even when using fewer data points. Labeling time and cost were also significantly lower; avoiding tagging by subject matter experts reduced costs by 18%.

A large global seed manufacturer partnered with Infosys to extract various data points from intellectual property documents related to studies and details of various experiments spread across geographies in different shared locations, languages, and versions.

SPEECH

Enormous conversational data floating in real time can be tapped to derive intelligence and improve business offerings. It could be faster customer issue resolution, product feedback acquisition, cost reduction, or workforce training.

Voice-based unstructured conversations have become the next big source of intelligence, presenting significant opportunities for enterprises. However, clean transcription of these conversations involves technical challenges such as different languages (English, Chinese, French, etc.), conversation-centric vocabulary, accents, ambient noise, and different channels (such as mono and stereo) used for recording conversations.

Over many years, large players such as Microsoft, Google, and IBM have gathered a huge corpus of voice data and created proprietary speech-based models promising high-quality output. However, these models have had privacy concerns due to cloud-based operations and limited customization options. Then, other open-source engines—including Kaldi, CMU Sphinx, Mozilla DeepSpeech, and Meta's wav2Letter—came up with better customization, but with different control levels, granularity of training data, effort requirements, and output accuracy. Even just a couple of years ago, open-source models were limited in terms of capabilities, but there has been a significant shift with these becoming increasingly sophisticated and proficient.

Trend 4: Speech processing through deep learning

In the past year, deep-learning models have taken over the majority of speech processing, replacing conventional models. These neural network models have substantially improved the quality of speech recognition, text-to-speech (TTS), speech diarization, among others. Some of the most popular ones are:

- Automatic speech recognition (ASR): wav2vec 2.0, Mozilla DeepSpeech, VoiceFilter-Lite (Google proprietary), Jasper, Quartznet.
- Diarization: Marblenet and Spearkernet.
- TTS:
- Spectrogram generation: Tacotron2, GlowTTS, FastSpeech2, FastPitch.
- Vocoders: WaveGlow, SqueezeWave, UniGlow, MelGAN, HiFiGAN.

This technology has led to significant advances in conversational intelligence, with applications such as knowledge mining, customer service, cross-sell and upsell marketing, and transactions across digital channels. Speech processing has also

recently removed the mandate that chatbots have no personality. Many systems, including Mozilla DeepSpeech and Infosys Nia, exhibit profound knowledge of many subject areas, mitigating scripting errors through continuous self-learning capabilities.

A global airplane manufacturer wanted to transcribe conversations between pilots and ground staff to boost operational efficiency. These conversations were studded with cockpit noise, strong regional accents, different languages, and heavy ambient noise. The company partnered with Infosys to develop a deep-learning open-source model that was custom-trained for accent variations. The model delivered high transcription accuracy, ran language insights to infer causes of flight landing delay and air accidents, and provided insights to improve ground staff and pilot training.

Trend 5: Open-source models now comparable to commercial counterparts

Traditionally, speech processing models, backed by large speech-to-text (STT) and TTS corpora, dominated the market. Most of these models, offered via cloud services, belonged to large tech giants. However, open-source models are advancing at speed. A majority of deep-learning models are open-source, primarily due to two factors. First, large transformers models meant for language processing are made available via websites such as HuggingFace and can run on machines with low computational power. Second, tech conglomerates such as Google, Microsoft, and NVIDIA have released some powerful proprietary models for the open-source community. This clearly indicates that open-source models will bring the next wave of transformation in speech processing models.

A large U.S.-based railroad company wanted to transcribe call center conversations to optimize operations, upskill the workforce, and improve customer satisfaction. The company partnered with Infosys to develop open-source custom models and framework. Using these technical calls, Infosys helped the railroad company transcribe audio files and perform text analytics to detect common reasons for calls. It also helped the company get better customer insights and identify workforce training requirements.

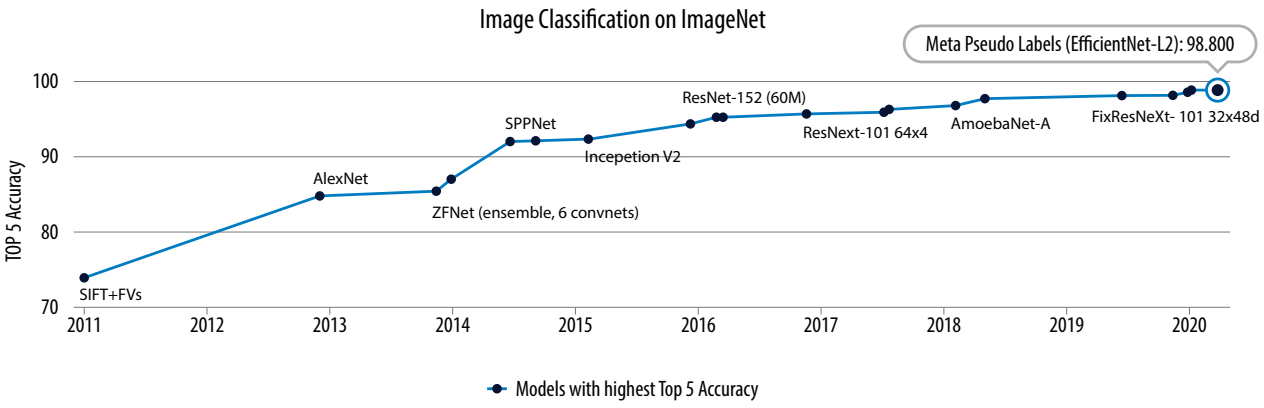
Trend 6: End-to-end conversational offerings in focus

Offerings that ease the deployment of speech processing with simultaneous services, such as STT, text synthesis, and TTS, are becoming widely available. With these prominent capabilities, businesses can deploy speech processing for multiple problems simultaneously and achieve faster results. Popular models include Mycroft, SpeechBrain, ESPNet, and NVIDIA NeMo. NeMo has separate collections for ASR, NLP, and TTS. Every module used in the pre-trained toolkit repository can be customized, composed, and extended to create new end-to-end conversational AI model architectures.

COMPUTER VISION

State-of-the-art pre-trained open-source models and neural network architectures have significantly improved computer vision (CV) implementations in organizations to solve image- and video-based insight problems. With the global ImageNet Large Scale Visual Recognition Challenge, the image classification error rate has decreased from >25% to <2% in the past 11 years, surpassing the human accuracy level. The continuous advent of new convolutional neural network architectures, such as AlexNet, VGG, GoogleNet, ResNet, and ResNeXt-50, to the most current EfficientNet-L2 that uses compound model scaling, with an error rate of less than 2% currently, has contributed significantly. These competitions prove that accuracy and performance do not solely depend on training data volumes, but neural net architectures play a differentiating role.[13]

Figure 4. Evolution of various state-of-the-art neural network architecture(s) from ImageNet competitions[14]



Source: Papers on Code

Trend 7: Image segmentation, classification, and attribute extraction through AI

Object detection, segmentation, and classification are the building blocks to address complex computer vision challenges. Object detection helps identify an object in the image, forms a rectangular boundary, and creates a bounding box to narrow down the object. Then, image segmentation identifies the object with all curves, lines, and the exact shape. This helps in a more granular and finer identification. This process also helps to establish various insights from images by classifying them, segmenting specific information, and/or extracting any image attributes. Object classification helps classify a particular object into a class or subclass — for example, vehicle classification by type (car, airplane, etc.), and then by brand (Audi, BMW, etc.).

This technique has considerably benefited the health care sector — identifying and narrowing tumor regions and further classifying them as malignant or not.

A large global energy company partnered with Infosys to optimize its cable diagnostics and repair operations to identify faulty cables based on the picture sent from the site, allowing them to take appropriate action. This helped the company to save costs and efforts.

A large global retailer wanted to develop a solution to extract and classify information from digitally scanned product art (SmartArt). The company partnered with Infosys to develop an AI model that can extract information accurately. The information could be further classified as contents, ingredients, instructions, etc. This made the information available on multiple channels for regulatory and compliance purposes.

Trend 8: AI and cloud power video insights

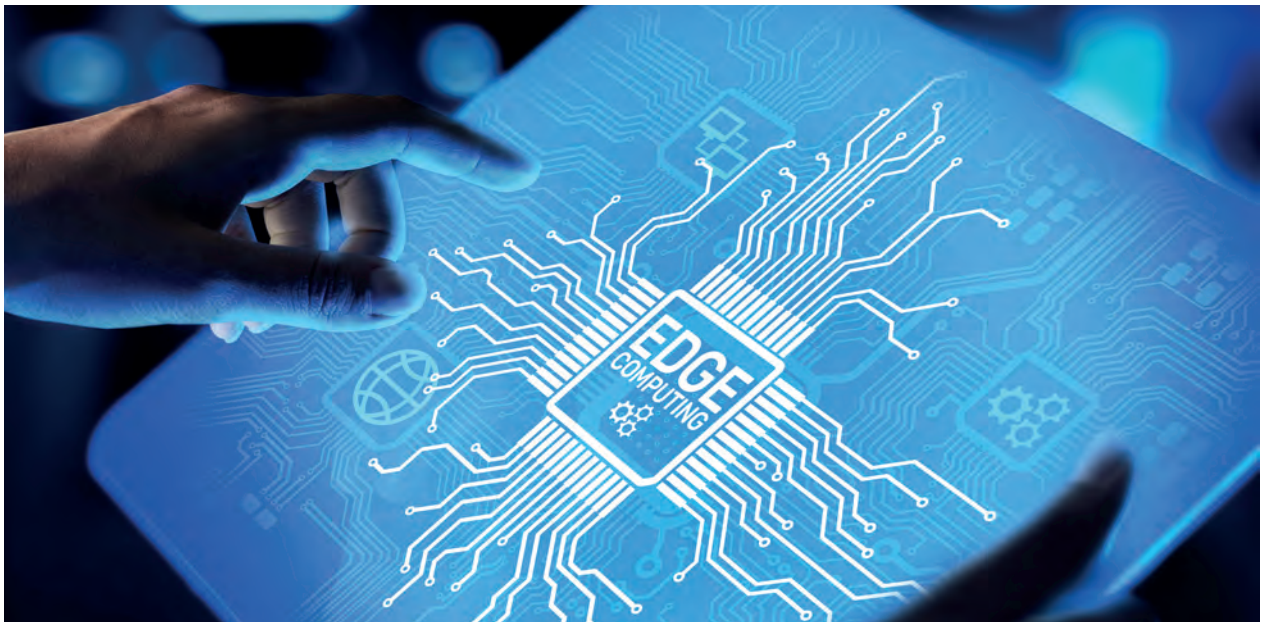
AI's application to videos offers interesting possibilities, such as generating video captions, video highlights, content moderation, brand coverage timings, surveillance, and people/object tracking. For applications like these, cloud computing is necessary for most inference tasks. In fact, object tracking and surveillance are far more powerful in the cloud than on devices, even with new advances in light detection and ranging technology on edge devices such as iPhone. For video processing, the cloud provides scene segmentation and multicamera scene reconstruction; on devices (or the edge — see the next section), only basic segmentation and de-noising capabilities are achieved.

As part of a prestigious global tennis tournament, Infosys extracted various game insights using CV-based algorithms. Event highlights, such as players waving to the crowd, extracting the score from the video feed, recognizing players, and determining the timeframe of a particular advertisement during the live telecast, were created.

Similarly, a large railroad company in the U.S. identified various assets spread across geographies with the help of a streaming video feed from a train-mounted camera.

Other use cases in the CV space include the digitization of KYC form filling, activity and pose recognition in videos, video synthesis, video summarization, and image captioning by leveraging state-of-the-art AI models and techniques such as 3D object detection, generative networks, and single-shot learning.

AI ON THE EDGE



Trend 9: Edge-based intelligence to address latency and point-specific contextual learning

Smart reply, auto suggestions for grammar, sentence completion while typing on a phone, voice recognition, voice assistants, facial biometrics to unlock a phone or an autonomous vehicle navigation system, robotics, augmented reality applications — all use local, natively deployed AI models to improve the response time. In the absence of a local AI model, the inference or prediction would be based on a remote server, and the experience would be suboptimal. Edge-based AI plays a quintessential role in remote locations, where network connectivity may not be continuous. Response times should be in the fractions of seconds and network latency cannot be afforded. Further, hypercontextualization is required with user-specific data.

Edge-based AI is feasible because of a significant improvement in edge processing-specialized embedded chip hardware and software such as Google tensor processing unit, field-programmable gate arrays, and GPU.

At the edge, two things happen — "inference/prediction" and "training/learning." For the inference or prediction to happen, a lightweight model is available. The model with this training capability can use local context-based learning and synchronize with the central model at the appropriate time. The synchronization can be done by just sharing the model parameters, weights, features, etc., without compromising on data privacy. Once the central model builds itself with several such updates from different remote edge-based AI models, it can update its training and share the updated model footprint with all the edge-based devices or clients, ensuring everybody gets the benefit of the central learning capacity. This process of distributed learning is called federated learning. This is employed as a strategy where sharing data has challenges of data privacy, shareability, network transport limitations, etc., but at the same time needs to leverage the benefits of abstracted learning through central capacity.[15]

TensorFlow Lite provides a complete toolkit to convert TensorFlow models to TensorFlow Lite, which can run on edge devices. Even with smaller models trained on less data, TensorFlow Lite gains the benefits of the central processing unit and GPU acceleration devices. MobileNet models convert several state-of-the-art convolutional neural network models to device models by sizing network architecture patterns such as depth-wise separable convolutions, hyperparameter optimization for width multipliers, and resolution multipliers with the corresponding trade-offs in accuracy and latency.[16]

A large telecom company, in collaboration with Infosys, developed a robust video intelligence solution for smart spaces. The solution takes data through real-time streaming protocol (RTSP) from CCTV, runs deep learning and computer vision models to detect humans across feeds, and tracks motion/movement. It derives insights such as people density in an area, ingress/egress count, dwell time analysis, and wait time analysis.

These models have been extended for social distancing compliance. Infosys developed an elevated body temperature detection solution as part of the COVID-19 response. The solution ingests feeds from thermal camera to edge devices. It then runs deep learning and computer vision models on edge to deduce a person's accurate temperature depicted within the thermal feed. Thereafter, it compares against the set threshold for detecting elevated body temperature as an indication of febrile symptoms.

Infosys has also built India's first commercial autonomous golf cart that uses an autonomous platform. Autonomous vehicles use deep learning and computer vision models to detect objects and lanes for autonomous navigation. It uses simultaneous localization and mapping (SLAM) and visual SLAM (vSLAM) models. The first autonomous vehicle is now deployed in a leading automotive OEM plant.

DATA ENGINEERING

Data engineering has been pivotal in accelerating analytic decision-making, operationalizing business value through ML Ops, and driving accurate decision-making on reliable data.

As much as 80% of an AI project involves data cleansing, preparation and data engineering activities, making data engineering crucial for sentience, and automation in data-centric services.

Trend 10: AI-powered technologies enhance data scientists' experience

Even today, many data scientists manually analyze data by using various techniques, with the need to apply various data cleansing activities. There is no standardized set of tools for data wrangling, analytics, feature engineering, and model experimentation. However, data scientists are increasingly shifting from an artisan to an industrialized ecosystem, leading to increased adoption of automated advisory for data cleansing and wrangling for faster feature engineering and quality analysis.

Legitimate privacy concerns, new regulations, cost pressure, and inherent data bias have pushed enterprises to explore data augmentation, an automated process for preparing data and synthesis.

One subset of this automated approach is synthetic data generation. Here, data is synthesized from scratch when no data is available, or when outliers/edge cases are rare in real-world data. This approach should be used when safe, reliable, fair, and inclusive ML models are required.

> A Europe-based telco wanted to use customer data to enhance client retention. The company worked with Infosys to build datasets to effectively predict customer churn. The telco reduced churn by 10%-15% by developing a catalog of customized offers.

Trend 11: Responsible data crucial for safe and sound AI development[17]

Explainable AI through responsible data is still evolving. The bias on data can have devastating effects on business outcomes, causing serious ethical and regulatory issues. The application of responsible and ethical data policies in AI development is beneficial for businesses and societies.

The rising dependability of AI applications on data to develop and train algorithms highlights the importance of secure and reliable systems.[18] Businesses must consider the following elements as a part of AI design principles:

- Identify data origin and data lineage.
- Identify the use of internal and public data for building models.
- Identify potential data corruption and anomaly detection.
- Protect individual data privacy rights.
- Resist cyberattacks.
- Comply with legal and regulatory requirements.

Trend 12: AI-based tools enhance data-quality

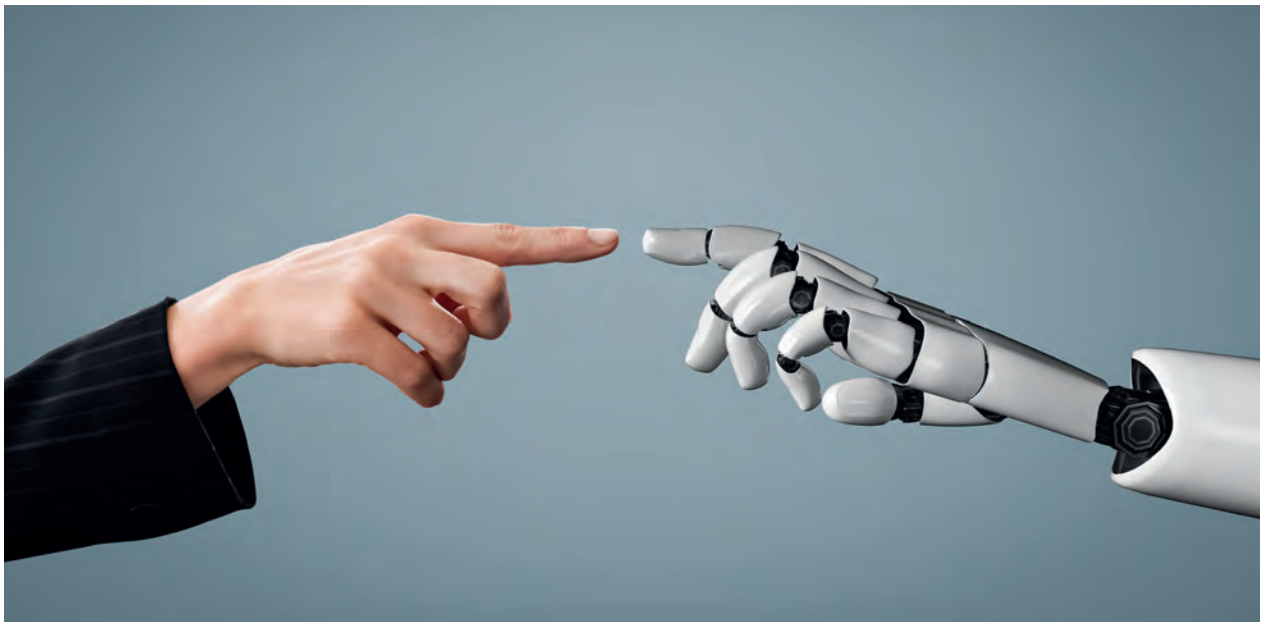
Whether it is for decision-making by corporate executives, frontline staff, or intelligent ML models, any intelligent enterprise needs high-quality data to operate. However, data quality issues are widespread. AI-based data-quality analysis has become an integral part of the ML Ops pipeline.

Enterprises have started considering data engineering an integral part of their data strategy. Tools such as Lakehouse, metadata management, data lineage, data quality, and data discovery will play a significant role in the data engineering architecture.

For data sharing between big enterprises, another technique of note is called "cooperative computing." This technique is deployed when robust datasets are needed for innovative new corporate ML models at scale and speed. In this paradigm, datasets are consolidated and encoded, facilitating different users to use these datasets efficiently and effectively.

An investment firm undergoing a modernization exercise wanted to build a data pipeline on AWS for corporate customers. It involved identifying, ingesting, cleansing, and loading the existing data in its legacy IT ecosystem built on mainframes. The firm partnered with Infosys to leverage Infosys Data Workbench to build a quality gate, where the data from mainframes could be profiled and ingested for building data-centric services. This included demand forecasting and identifying the next best action. The customer improved the marketing campaign effectiveness by 70%, with effort savings of around 45% for the commercial sales line.

RESPONSIBLE AI



With the increasing adoption of AI systems in critical decision-making, monitoring the credibility of associated outcomes becomes even more critical. There have been instances of erroneous AI-driven outcomes that led to ethical implications. For instance, an AI hiring algorithm was found biased against specific races and gender, and an AI algorithm used by a car insurance company by default classified males under 25 years old as "reckless" drivers. Further, AI algorithms used to derive recidivism score, used by judges to make decisions, were found biased against black defendants.

This attracted negative fame, lawsuits, and other societal implications to the point that regulators, official bodies, and general users now seek more transparency in AI-driven decisions. In the U.S., insurance companies need to explain their rates and coverage decisions, while the European Union introduced the right to explanation in the 2018 General Data Protection Regulation.

Trend 13: AI ethics throughout the development lifecycle

Responsible AI concepts should be factored in from the beginning to ensure the business stays out of any AI ethics and bias issues. Explainability is one such critical concept. The design and development teams should be aware and informed of every step in the AI lifecycle to answer any related questions, providing all information AI users would seek to understand how and why the system made a decision. This way, an organization can remain clear of adverse ethical issues and maintain customer trust.

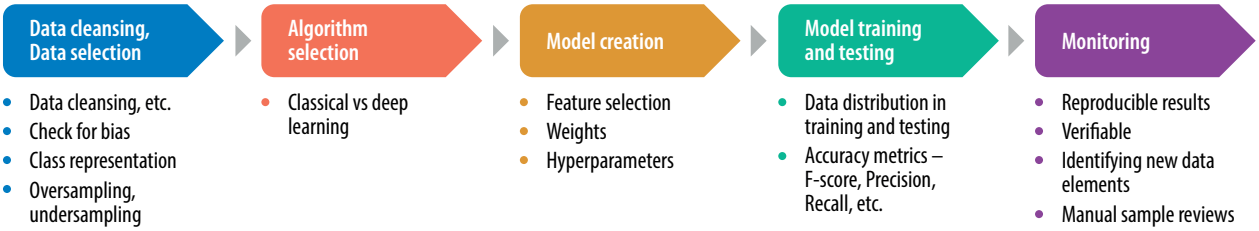
These scenarios demand efficient tools to make AI systems more transparent and interpretable, ensuring trust, fairness, transparency, reliability, and auditability. AI models should adhere to the following principles:

- **Purposeful:** An AI system should be designed with empathy and follow a human-centric approach with socially responsible use cases. For example, consider user preferences and behavior to provide recommendations.
- **Ethical:** Models should comply with legal and social structures and be designed with high-cost functions that prevent unethical behavior. There should be transparency in data and models.
- **Human reviewed:** Although AI models are built to operate independently without human interference, human dependency is a necessity in some cases. For example, in fraud detection or cases where law enforcement is involved, human supervision is required to review decisions made by AI models.

- **Bias detection:** An unbiased dataset is an important prerequisite for reliable and nondiscriminatory predictions. AI models are being used for credit scoring by banks, resume shortlisting, and in some judicial systems. However, some datasets were found with an inherent bias toward color, age, and/or sex.
- **Explainable:** Models should enable easy interpretation of results such as predictions, recommendations, etc. Explainable AI helps understand the decision-making process of AI systems and recognize which features of the given input are emphasized while making predictions.
- **Accountable:** Models should use telemetry for auditing all human and machine actions. There should be data lineage for traceability, and all models/datasets should be version controlled.
- **Reproductive:** The ML model should be consistent when giving predictions. Many practitioners think that explainable AI (XAI) is applied only at the output stage, but the role of XAI is throughout the whole AI lifecycle.[19]

Thus, consistent and continuous governance can make AI systems understandable and resilient in various situations.

Figure 5. XAI in the AI lifecycle



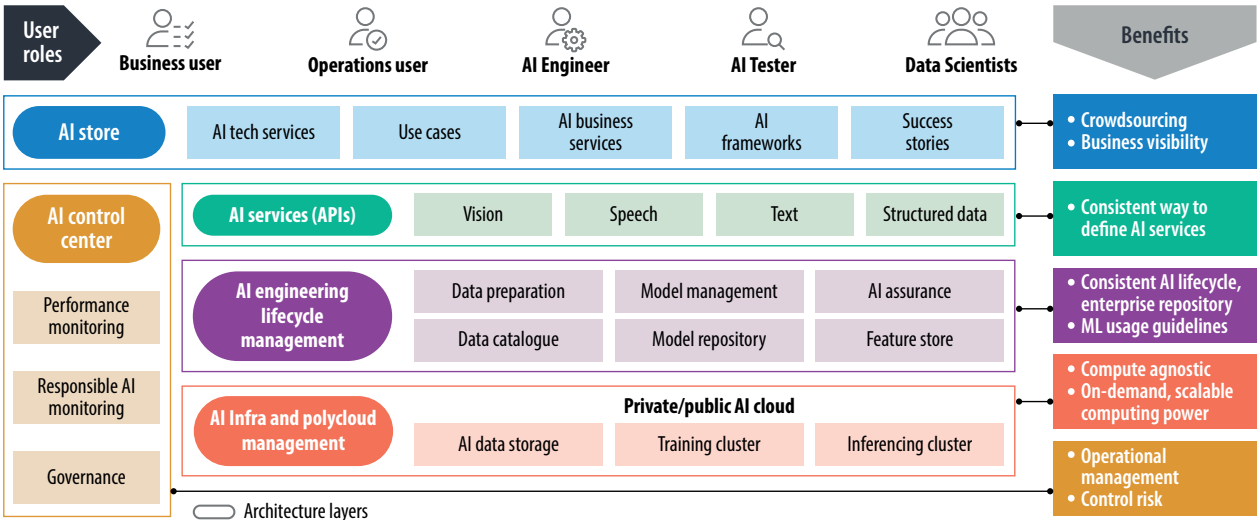
Source: Infosys

AI PLATFORMS

Growing AI pervasiveness has prompted enterprises to move from a function-specific approach to a platform-based AI approach, ensuring unified development focused on organizational needs and goals. Infosys proposes a modular reference architecture that can help organizations keep AI future safe and swiftly adapt to future developments.[20] The platform keeps AI future-proof by ensuring the following:

- AI development through a unified organization-level vision.
- Least dependency on specific technologies and vendors.
- Best-in-class solutions based on available technology.

Figure 6. Reference architecture for an enterprise AI platform



Source: Infosys

An AI platform approach enables AI stakeholders to participate and engage cross-functionally in the development process. Infosys research shows that the inability to work across AI silos is a big pain point for AI initiatives; this platform-centric approach is vital to scale and institutionalized AI. Visionaries in this research used platforms to bring disparate units together in cross-functional teams. They fostered superior AI governance approaches and delivered better operations by reimagining business processes with AI at the core.

Trend 14: Integrated AI lifecycle tools to drive industrialized AI[21]

Enterprises cannot afford to take an artisan approach to AI and experiment with pilots and a handful of disparate AI systems built in silos. Without focusing on achieving AI at scale, data scientists created "shadow" IT environments on their laptops, using their preferred tools to fashion custom models from scratch and prepare data differently for each model.

The AI lifecycle involves various stages, from data collection, data analysis, feature engineering, and algorithm selection to model building, tuning, testing, deployment, management, monitoring, and feedback loops.

Infosys proposes a simplified ML to prepare data, develop the data model, deploy, and optimize models in production.

Optimizing AI model performance is an emerging area in the AI lifecycle. Intel provides toolsets to introspect model and measure performance, optimize math libraries and algorithms to improve model performance, and optimize versions of popular serving engines.

Based on client interactions, we are seeing a rise in the adoption of end-to-end AI lifecycle development tools, including H2O.ai, Kubeflow, and MLflow. However, the standardization of these tools and pipelines is still under progress.

A U.S.-based telecom player struggled with AI development, as it used multiple tools and had limited data for model development. The company partnered with Infosys to develop an industrialized AI system to build a pipeline for AI and ML developments over AWS. The telco was able to improve returns on its AI and ML investments.

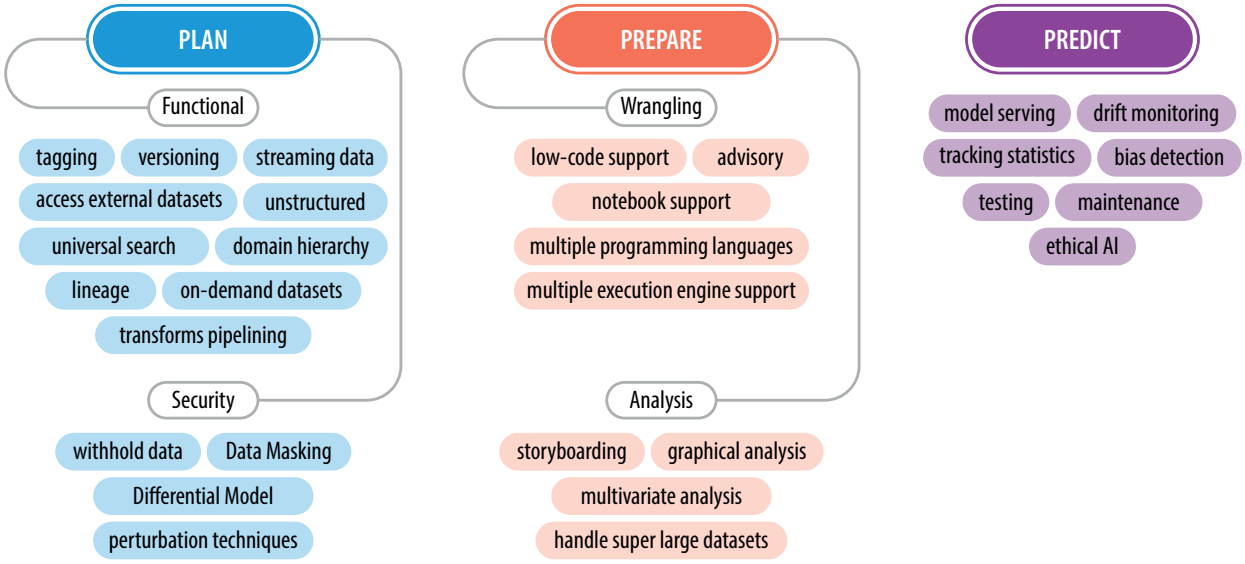
Trend 15: From data scientist to data engineer with automated ML

Data scientists spend around 80% of their efforts on finding data rather than building AI models. Creating an AI model from scratch needs effort and investment for collecting datasets, labeling data, choosing algorithms, defining network architecture, establishing hyperparameters, etc. Further, the choice of language, frameworks, libraries, client preferences, etc., differs from one AI problem to another.

Specialized roles such as data engineer and ML engineer offer skills vital for achieving scale. With the help of a rapidly expanding stack of technologies and services, teams have moved from a manual and development-focused approach to an automated, modular, and fit-to-address approach, from managing incoming data to monitoring and fixing live applications.

A global investment firm's data scientists were struggling to get the right data experience. The firm, in partnership with Infosys, built an efficient solution to plan, prepare, and predict data. The solution reduced the time spent on identifying the right tools and data, boosting data scientists' efficiency by up to 65%.

Figure 7. Components and capabilities of data box



Source: Infosys

References

Trend 1: Deep neural network architectures help improve generalization and accuracy

1 Advanced trends in AI: The Infosys way, October 2020, Infosys

Trend 2: Transition from system 1 to system 2 deep learning

2 System 2 deep learning: The next step toward artificial general intelligence, Dec. 23, 2019, TechTalks

Natural language processing

3 Attention Is All You Need, arXiv.org

4 Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing, Nov. 2, 2018, Google Blog

5 Better Language Models and Their Implications, Feb. 14, 2019, OpenAI

6 SuperGLUE

7 Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer, Feb. 24, 2020, Google Blog

8 OpenAI API, Sept. 18, 2020, OpenAI

9 Pretrained Models, July 12, 2021, GitHub

10 Scaling Language Model Training to a Trillion Parameters Using Megatron, April 12, 2021, NVIDIA

11 Your AI pair programmer, GitHub

12 OpenAI Codex, Aug. 10, 2021, OpenAI

Computer vision

13 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), ImageNet; ImageNet Object Localization Challenge, Kaggle; Image Classification on ImageNet, Papers with Code; Fixing the train-test resolution discrepancy: FixEfficientNetv, April 20, 2020, arXiv.org

14 Image Classification on ImageNet, Papers with Code

Trend 9: Edge-based intelligence to address latency and point-specific contextual learning

15 Convergence of Edge Computing and Deep Learning: A Comprehensive Survey, Jan. 28, 2020, arXiv.org

16 MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, April 17, 2017, arXiv.org

Trend 11: Responsible data crucial for safe and sound AI development

17 The Hidden Dangers of Data Science, Dec. 12, 2019, Towards Data Science

18 Implementing Responsible AI A Framework for the Next Normal, June 2021, Infosys Consulting Insights

Trend 13: AI ethics throughout the development lifecycle

19 The Need for Explainable AI, October 2020, Infosys Insights

AI Platforms

20 An Architecture for Mature Enterprise AI, November 2021, Infosys Knowledge Institute

Trend 14: Integrated AI lifecycle tools to drive industrialized AI

21 Scaling AI like a tech native: The CEO's role, Oct. 13, 2021, McKinsey

Advisory Council

Mohammed Rafee Tarafdar
SVP and Unit Technology Officer

Prasad Joshi
SVP, Emerging Technology Solutions

Balakrishna DR
SVP, Service Offering Head – Energy, Utilities, Communications & Services

Satish HC
EVP, Head Global Services – Data and Analytics

Dinesh HR
EVP, Head Global Services – Enterprise Package Application Services

John Gikopoulos
Partner – AI & Automation Practice

Shyam Kumar Doddavula
AVP, Emerging Technology Solutions

Thirumala A
VP – Education, Training and Assessment

Hasit Trivedi
AVP – AI & Automation

Venkata Seshu G (Seshu)
AVP – Data and Analytics

Contributors

Allahbaksh Mohammedali Asadullah

Amit Gaonkar

Chandra Sekhar Y

Dr Ravi Kumar G. V. V

Dr. Puranjoy Bhattacharya

Eggonu Vengal Reddy

Gopinadh Bapatla

Harleen Bedi

Jagadamba Krovvidi

Kamalkumar Rathinasamy

Karthik Andhiyur Nagarajan

Shyam Kumar Doddavula

Surya Prakash G.

Swaminathan Natarajan

Venkata Lakshminarayana Indraganti

Vijayaraghavan Varadharajan

Vittal Setty

Producers

Ramesh N
Infosys Knowledge Institute
ramesh_n03@infosys.com

Abhinav Shrivastava
Infosys Knowledge Institute
abhinav.s08@infosys.com

About Infosys Knowledge Institute

The Infosys Knowledge Institute helps industry leaders develop a deeper understanding of business and technology trends through compelling thought leadership. Our researchers and subject matter experts provide a fact base that aids decision-making on critical business and technology issues.

To view our research, visit Infosys Knowledge Institute at infosys.com/IKI or email us at iki@infosys.com.

For more information, contact askus@infosys.com