

TECH NAVIGATOR: THE AI-FIRST ORGANIZATION

Infosys®
Navigate your next



Infosys® | Knowledge Institute

THE INFOSYS TOPAZ EDITION

Table of contents

Executive summary	4
Explore the full report	7
Introduction	7
Building block 1: Reimagining experiences and processes	8
Building block 2: AI engineering excellence	11
Building block 3: Responsible AI by design	14
Building block 4: The AI operating model	17
Infosys Topaz – Accelerating AI-first business value for Infosys and our clients	20
References	21
Contributors	22
Authors	22
Acknowledgements	22

Executive summary



We are at the beginning of a major technological era, one dominated by artificial intelligence (AI). Firms that build on their digital and cloud investments with AI will leapfrog those that don't and will dominate industries and professions.

This Tech Navigator sets out to explore what savvy firms need to do to put those investments to work. While they need to be human-centric, as we outlined in [last year's report](#),¹ they will need to use AI to augment and amplify human potential, to become more innovative, to unlock efficiencies at scale, to grow faster, and to build a connected ecosystem.

Consumerization of AI

Generative AI is the latest wave in AI advancement. It follows from machine learning and predictive analytics, then deep learning, and now transformer architecture and foundation models. It is described by Stanford as "[models trained on broad](#)

[data, generally using self-supervision at scale, that can be adapted to a wide range of downstream tasks](#),²" and now used in the latest consumer products such as ChatGPT, which is built on OpenAI's GPT-4 large language model (LLM).

By leveraging the power of this general purpose AI technology, we are seeing enterprises evolve from using AI to manage operations and specific business functions to using it to reimagine the way customer experience and services are delivered by the business (Figure 1).

Becoming AI-first: The building blocks

AI-first firms will have a strategy in place for deploying these nascent models. They will understand which experiences and processes to amplify through AI, the tooling and automation needed to deploy these models, and will have the right talent and operating model to bring this AI to life.

Figure 1. Businesses should move across the three horizons to evolve as AI-first

KEY PATTERNS

- Billion/trillion parameter models
- Zero-shot learning
- Multitask learning
- Multimodal and multilingual
- Closed and open access models
- Responsible by design
- Auto ML

- AI governance – AI ethics, explainable AI
- Model pruning, quantization tech
- Transfer learning
- Neural networks
- Object detection, classification, segmentation

- Prediction recommendations
- Logistic regression
- Classification regression
- Rule-based
- Expression-based



H3
Transformer architectures, foundation models, generative AI (self-supervised)

AI models should be capable of learning and evolving on their own with minimal human intervention.

H2
Transfer learning, responsible AI (less data, explainable systems)

These are rapidly gaining prominence among enterprises. Businesses are investing in AI systems that are capable of making fast, transparent, and unbiased decisions.

H1
Conventional AI (augmenting intelligence)

These systems are already mainstream, providing AI-powered assistance to business decisions.

Source: Infosys



Throughout the report, we refer to the three waves of AI, categorized into Horizons 1, 2, and 3, or H1, H2, and H3. Horizons are a way to evaluate tech trends.³ Horizon 1, or H1 technologies, are well established and widely used. H2 is technologies that are in use and most of the ongoing work is happening using these technologies. H3 is emerging technologies that are used in pockets or for innovation pilots and include disruptive ideas across enterprises, but some have the potential to become mainstream. Advances in H3 can also create new risks in compliance, safety, and other areas that must be managed.

This is what we explore in the coming pages — a prescriptive framework composed of four building blocks (Figure 2) that we are using in Infosys’ own transformation from a cloud-native to now an AI-native enterprise. The four building blocks are:

- **AI-first experiences and processes** — to evaluate and identify experiences and processes that will benefit most from AI-first reimagination, including AI assistants.
- **AI engineering excellence** — to build foundational data and AI engineering processes, tools and automation in a bid to craft digital rails for delivering and scaling AI projects.
- **Responsible AI by design** — to create guardrails, controls and processes to ensure that all the AI products and services are trustworthy and meet regulatory guidelines and policies.

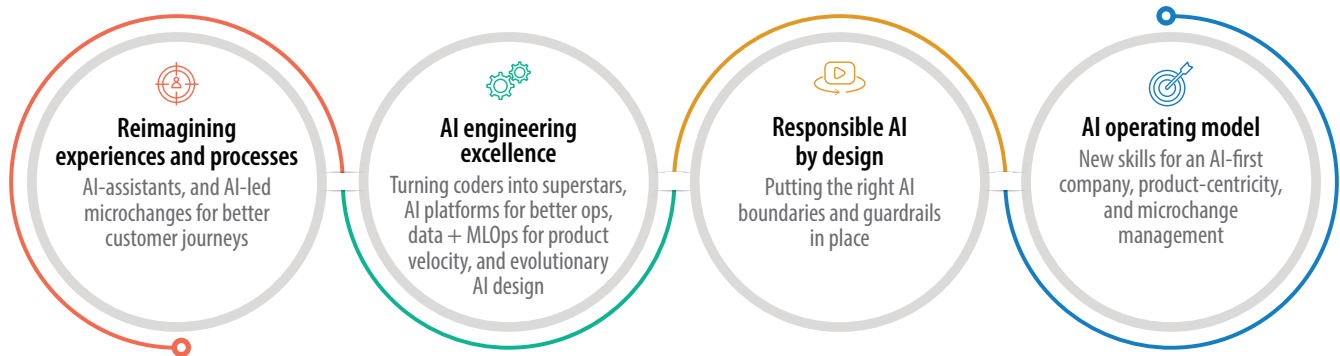
- **The AI operating model** — to build AI-first talent, processes and a product-centric operating model to design and deliver AI services.

Reimagining experiences and processes

Firms must prioritize the products, processes, and features that are most ripe for transformation through AI, and evaluate them in terms of business impact, ease of implementation, and trustworthiness.

An AI canvas can also be drawn up that covers business problems, expected end-user value, and the necessary guardrails and controls. Each experience must have a human-in-the-loop (HIL) and telemetry to improve AI model performance over time.

Figure 2. The four building blocks of an AI-first organization



These four building blocks create an enterprise that

- Amplifies human potential
- Unlocks access to intelligence embedded in artifacts, systems, and humans
- Drives exponential change for scaled impact

Source: Infosys

AI engineering excellence

In order to build good AI, the right engineering foundation and tooling design approaches are needed. A PolyAI approach ensures flexibility to choose the best-fit AI solution for a given problem, while making use of both open and closed AI models is needed depending on the use case.

Closed models such as GPT-4, from Open AI, can be used with generalized use cases to quickly realize value.

For specialized use cases, models, and IP, open models such as BLOOM and CodeGen, can be fine-tuned using a narrow transformer approach to create differentiation and competitive advantage. The entire software engineering and operations processes should also be augmented through AI assistants that can help improve the productivity of developers, testers, and operations teams.

Responsible AI by design

Third, firms should implement the idea of shift-left with responsible AI by design. This isn't easy, however. Although the use of AI is exploding, ethics and governance are scrambling to keep up, with some jurisdictions banning the use of generative models until better regulation is in place.

To ensure failsafe AI systems, we recommend baking ethics into every step of the AI engineering and application life cycle.

Our AI framework comprises five building blocks, starting with the objective of building a trustworthy process or product through appropriate governance, and then monitoring and

measuring progress, building capabilities, and making sure the AI is compliant with data protection, record keeping, and reporting.

AI operating model

In our [Digital Radar 2023](#),⁴ released earlier this year, we found that more important than the introduction of technology is the way the organization is set up to take advantage of it.

Going AI-first means recognizing that some jobs will be displaced and new roles like prompt engineers and model-tuners will be created in their place.

We recommend that organizations use a product-centric approach for both AI product development and core engineering. A product-centric mindset will therefore be vitally important, as will what we term “micro is the new mega” — this is a way of turning change projects into a series of micro-sprints that produce exponential results and business outcomes.

AI is not just another technology, but one that will upend the way that organizations make money in the future and remain competitive. No wonder then that according to [Stanford's AI Index report 2023](#)⁵ and CB Insights’[“State of Generative AI”](#)⁶ report, we are witnessing an explosion in AI development, with more than 37 LLMs available, \$2.6 billion in equity funding and 500,000 AI publications released in 2022 alone.

Are you on board? Read on to discover the perspectives on these technologies from Infosys experts, and what it means for your business.

Explore the full report



Introduction

No firm can afford to ignore AI, and many are already deploying it in some way. In our [2022 Data + AI Radar](#)⁷ report, we found that more than 80% of the 2,500 practitioners surveyed had deployed their first AI model into production within the last five years (Figure 3).

Every business will have to become AI-first. At Infosys, we are amid a strategic program to become an AI-first company, based on three key objectives:

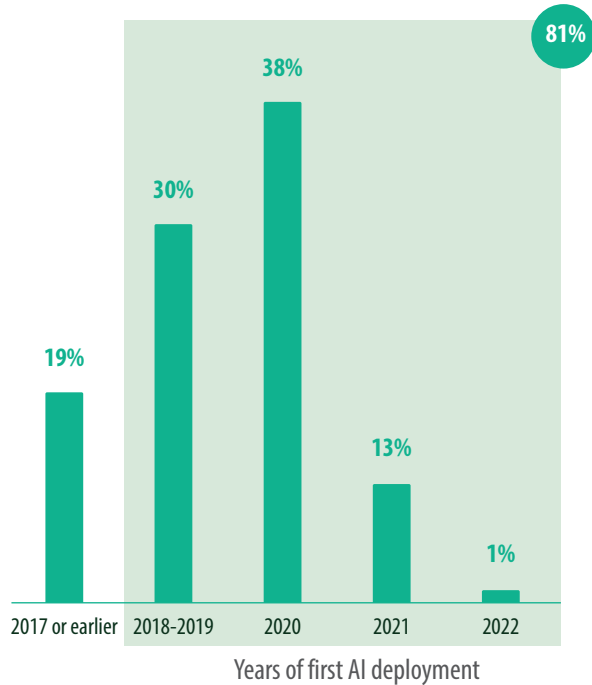
- Amplify the potential of our workforce of more than 350,000 employees.
- Unlock value from our understanding of industries and our nearly 2,000 client business and IT landscapes.
- Create exponential impact.

Infosys has also launched Topaz, an AI-first set of services, solutions, and platforms using generative AI technologies that will accelerate this metamorphosis, while delivering significant value to our clients as well.

At the heart of this transformation is the explosion of generative AI use cases, based on large language models and foundation models.

Foundation models are typically self-supervised (not requiring labeled data), large (billions of parameters), pre-trained, generalized (not specific to a task) and adaptable (through prompts). They can be fine-tuned for specialized domains and tasks and improved with continuous human feedback. Large language models such as GPT-4 are a type of foundation model that have been used to build consumer AI products.

Figure 3. Four out of five companies deployed their first AI model less than five years ago



Source: Infosys

These models have led to an explosion of generative AI tools that can create text, code, images, video, audio and 3D models. We are also seeing the emergence of multi-modal models that can operate across a number of mediums such as text and images, and multi-language models that can understand and operate across several languages. All these are available either as open models or closed models.

At Infosys, we view AI in two broad categories: core AI engineering and applied AI.

Core AI engineering focuses on the identification, fine-tuning, and deployment of models, application programming interfaces (APIs), and platforms in a responsible manner. Applied AI uses these services to build products, from demand planning and code refactoring to document analysis and helpdesk self-service.

As we discuss in this Tech Navigator, businesses that aim to become AI-first need to reimagine experiences and processes, operating models, and talent acquisition.

They should establish a robust responsible by design framework for ethical AI usage.

In short, an AI-first organization puts AI at the heart of everything it does, from hiring talent to delivering products and services.

Building block 1: Reimagining experiences and processes

Birth of the AI assistant

As we posited in last year's *Tech Navigator - Building the Human-centric Future*,⁸ technology has the potential to empower individuals. AI-first organizations will offer experiences that unlock the full potential of their workforce.

Here, technologies augment rather than replace human capabilities. One way to do this is to use AI assistants, based on generative AI. The vision is that AI assistants will help individuals streamline their work by automating tasks such as code-writing, data review, and document classification. This means humans are released from time-consuming tasks and can turn their attention to value-generating work.

Deriving business value from this technology requires firms to think of the personas, experiences, processes, and tasks that can be reimaged using AI and empowered through AI assistants. "As we go AI-first, we'll be giving each Infosys employee a personalized AI assistant," says Nandan Nilekani, chairman of Infosys. "Through generative AI capabilities, the AI assistant will be a steady and empathetic human counterpart, well versed in the day-to-day activities that employees have to do, while amplifying their creative and human potential in the process."










As we go AI-first, we'll be giving each Infosys employee an AI assistant. Through generative AI capabilities, the assistant will be a steady and empathetic human counterpart, well versed in the day-to-day activities that employees have to do, while amplifying their creative and human potential in the process.

Nandan Nilekani
Chairman of Infosys



Figure 4. Applying AI assistants to the software engineering life cycle

	Life cycle stage	AI augmentation through AI assistants
1	 Project planning and analysis	<ul style="list-style-type: none"> • Effort estimation planning and analysis • Risk assessment • Instant simulation
2	 User stories and backlog development	<ul style="list-style-type: none"> • Document completion and suggestion • Product features completion • Requirements completion analysis • Knowledge management
3	 Visual and technical design	<ul style="list-style-type: none"> • Image generation, • Image inpainting • Headline/copy texts • Website and code generation
4	 Build	<ul style="list-style-type: none"> • Code generation • Code completion • Code documentation • Code translation • Design pattern implementation • Unit test case generation • Bug prediction • Application security testing
5	 Test	<ul style="list-style-type: none"> • Optimize the number and value of tests • Eliminate redundant tests • Automated test script generation, self-healing of scripts, visual regression, test suite optimization, defect prediction, automated test selection – based on code changes, coverage analysis
6	 Operate	<ul style="list-style-type: none"> • AIOps • Predictive failures and actions • Digital workers (orchestrating tasks) • Knowledge management
7	 Refactor	<ul style="list-style-type: none"> • Code refactoring • Bug predictions • Application security testing

Source: Infosys

Figure 4 sets out how these AI assistants can be applied to the software engineering life cycle. These AI assistants will help organizations that have embraced AI-first to maximize profit from improved operations (what we term “enterprise left brain”) while at the same time creating new forms of value (“enterprise right brain”).

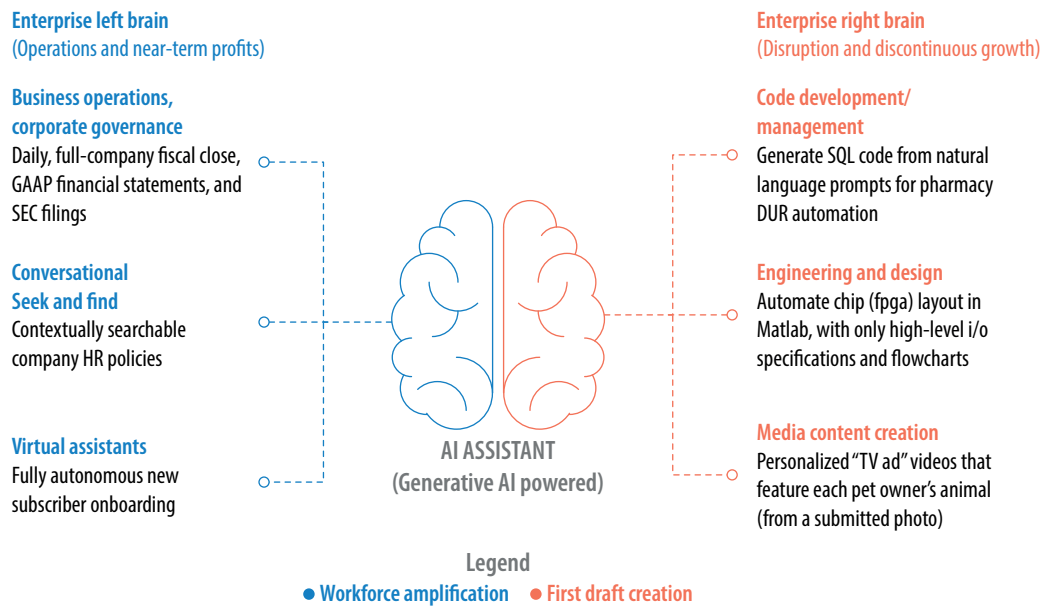
We describe this AI-first organization as “structurally ambidextrous” (Figure 5).

All AI-assistant development initiatives should be evaluated and compared along three fundamental dimensions: business impact, ease of implementation, and trustworthiness.

The AI assistant must also have explainability, governance, and provenance baked into the design of the LLM, a theme we turn to in the third section of this report.

AI assistants are an example of AI-in-the-flow, where

Figure 5. Structural ambidexterity in corporate innovation



Source: Infosys

individuals can do their work in one screen, with underlying technologies abstracted away. We believe that this AI-augmented user experience will improve employee satisfaction and, as [Digital Radar 2023 found](#),⁹ increase employee retention.

AI-powered processes

AI enables better, faster, and more automated processes. Better processes lead to better outcomes, which enhance efficiencies.

However, the change initiative must be considered carefully. Some firms plug AI into existing processes to make employees’ lives easier or to generate ad-hoc customer insights. This has little impact on overall organization health. Conversely, others attempt to overhaul the entire organization at once — but are overwhelmed by too many moving parts, stakeholders, and the sheer number of AI initiatives.

An AI-first firm is structurally ambidextrous, good at both operations and creating new forms of value

One approach is to first identify processes that benefit from AI-led microchanges, then to use AI assets to deliver the reimagined process. The Infosys AI Store provides more than 12,000 AI use cases, including generative AI use cases and more than 150 pretrained AI models, 100 datasets, and 50 AI templates to unlock process value at scale.

We also recommend developing an AI canvas for each of the prioritized AI processes. This covers the business problem, expected business value, expected end-user value, data strategy for training and modelling, an objective function for measuring effectiveness, and the guardrails and controls to be put in place during implementation.

Some industries are ripe for business process redesign, while many are advanced in their journey. Insurance companies, for example, use advanced AI to improve client onboarding and underwriting. Utilities have begun to use AI for equipment maintenance processes and procedures, especially for remote areas.

Computer vision algorithms are automating home and car repair, while AI-based telemedicine is changing the way healthcare is delivered.

Mercedes-Benz aims to use advanced AI to shorten the time its cars spend on the test bench. The impact goes beyond faster test cycles to also decreased CO2 emissions without reducing standards. At Infosys, we are adding AI assistants to its mobile employee experience layer for sales personnel to incorporate fresh relevant market intelligence, as part of our AI-first sales tool operations.

The caveat is that no matter how much AI is used in process re-engineering, a human in the loop is always necessary, to comply with regulations and to ensure trust, transparency, and explainability. Telemetry should be used to capture feedback

on AI effectiveness and use the data to improve the AI model performance over time.

Process re-engineering is reaching another level of performance in the AI era. Smart companies view the introduction of AI as rationale for a fresh perspective and higher expectations for end-to-end processes and customer journeys. As firms increase their use of generative AI, they will automate or augment everyday tasks and reimagine their business processes. Entirely [new business models](#)¹⁰ will emerge to generate revenue, and the operating models will follow to make them a reality. AI-first business models and experiences will then allow small businesses to appear big and incumbents to move faster.

Building block 2: AI engineering excellence

Develop and nurture top-tier coders

Coders are in short supply and overworked, and they suffer from burn-out and reduced productivity and efficiency. New tools are helping them cope with burnout and adding value to the organizations that deploy them.

Microsoft is an AI-first business investing heavily in this area, having recently committed [\\$10 billion to research and development at OpenAI](#),¹¹ solidifying its position as a leader in the AI industry.

Tools such as OpenAI's Codex, GitHub's Copilot and Open AI's ChatGPT can be used to complete lines of code and to find bugs. They also create code in most programming languages from a natural language prompt. AI-based techniques assist in other stages of the software development life cycle, including gathering requirements, design, deployment, and maintenance.

These AI tools, based on large language models, can perform a variety of tasks but demand heavy compute for inferencing. Infosys addresses this issue by building narrow transformers. Here, an appropriate smaller language model is used as the foundation model, and then fine-tuned with domain and task-specific data to augment a specialized task, such as code completion.

Full fine-tuning updates all parameters of the foundation model, demanding higher compute, and when there are multiple downstream tasks, full fine-tuning results in a new model for each task demanding more storage.

Infosys adopts parameter-efficient fine-tuning (PEFT) methods

to build narrow transformers as these methods enable adaptation of foundation models to downstream tasks with fine-tuning only a small number of (extra) model parameters. This results in reduced computational and storage costs, while achieving performance comparable to that of full fine-tuning. These narrow transformers are built and served at scale without external compute, ensuring data security.

In the [DevSecOps](#)¹² age, developers are not just responsible for writing code, but also for setting up the production environment, and assembling, integrating, customizing, and maintaining applications once in production. This burden is now carried by another 2023 software tech trend — [platform engineering](#).¹³ This accelerates development velocity and reduces software engineering overload through self-service capabilities — standardized tools, components, and automated processes — offered via an internal development platform (IDP). Gartner expects that by 2026, [80% of organizations will establish these platform teams as internal providers of reusable services, components, and tools for application delivery](#).¹⁴ Beyond AI-driven coding, these platforms encourage consistency and efficiency in software development and provide relief from the management of delivery pipelines and low-level infrastructure.

As AI tools evolve, the software development sector is in for an operational and financial windfall. If firms can motivate their workforce to upskill in AI prompt engineering, the future will be an AI-first, continuously learning and evolving organization — an AI-first [Live Enterprise](#).¹⁵

AI-first for better operations

In the first AI wave (H1), systems and methods were predominantly used to supervise and optimize operations. In the AI-first era, these methods will pave the way to further improve operations and increase efficiency. We will see the onset of newer frontline ML methods such as reinforcement learning, backed by meta-learning-driven dynamic control.

Our paper on a [self-driving cloud for greener business](#)¹⁶ discusses this trend, where companies analyze data using ML techniques to make informed decisions and take proactive measures. The dynamic control feature also means the system can learn and adapt to new scenarios on the fly, driving even more efficient operations.

AI-first operations improve customer and employee experience through generative AI. Using AI assistants, an AI-first organization can handle internal and external user queries at pace, improving productivity and innovation — and then they also become companies that people want to work for. Data-driven organizations that care about post-sales customer engagement improve employee retention, as the

research for our [Digital Radar 2023](#) report reveals.¹⁷ These AI systems use machine translation to interact with users, driving localized contextual conversations that add further data points and improve efficiencies and insights.

Many applications are possible, including voice assistants and recommendation systems.

Developers and organizations can build applications and services that understand and interact with users in more natural and intuitive ways.

AI-first customer-centric operations systems deliver the following capabilities:

- Offer multilingual support.
- Improve efficiencies through task automation and FAQ, reducing reliance on human professionals.
- Customize responses based on the learnings from customer interactions.
- Increase UX and CX through operations efficiency for CRM and ERP integrations.
- Enhance operational scalability by reducing the necessity for additional customer interactions, such as during Samsung/iPhone launches in the retail and technology sectors.

This technology is also well suited to solve problems across industries such as telecommunications, utilities, and retail. Google's Dialogflow is used by Optus, one of the largest telcos in Australia, to power virtual agents in a support application. Because the technology comes with prebuilt agents, in-depth programming knowledge is not required, and the technology can be rolled out much more quickly.

For example, prebuilt agents answer requests such as "I need help paying my bills" or "I haven't received my order, where is it?" without requiring custom programming.

These capabilities extend to other industries too. Healthcare providers can extend health bot instances to include novel scenarios and integrate them with other IT systems and data sources.

AI platforms enable operations with self-heal capabilities, anomaly detection, automated monitoring, and alerting. For example, predictive maintenance is a great use case in the utilities sector.

The AI organizations of the future will need to respond in close to real time to queries across platforms, including mobile, web, chatbots, smart devices, interactive voice response systems, and messaging apps.

AI-first operations, using advances in NLP and transfer learning,

are the future of time-limited, data-driven conversations, extending from internal support to customer contact centers.

Data and MLOps to drive velocity

In H2, data engineering was key: in our 2021 paper [Scaling AI: Data over Models](#),¹⁸ we estimated that 25% to 60% of machine learning project costs at that time were spent on manual data labeling and validation.

Firms had to manage data lineage and build systems with active learning, in which a classifier examines unlabeled data and selects part of this data for further human labeling. For the process to operate effectively, machine learning systems needed to be efficient, scalable, and reliable. This landscape also required a central model repository and trustworthy AI practices.

Many firms are still working in H2, and technologies such as Azure ML, AWS SageMaker, MLFlow, and products such as DataRobot and Iguazio, are emerging as sources for model management, deployment, and managing training data. Meanwhile, our clients require online and offline feature storage for machine learning data management and monitoring.

We are now in the H3 era, and all eyes are on generative AI and the models that underpin these tools. Building these models is a complex process. It can take large firms several hundred days and thousands of CPUs and GPUs to birth a new large language model.

Creators of these models now rely on MLOps to support scaling techniques, including data parallelism, pipeline parallelism, and tensor model parallelism.

With data parallelism, tasks are run in parallel, with data divided into partitions, and the models run on separate subsets of the data, increasing model training speed. Model parallelism, as the name suggests, divides a massive complex model either vertically or horizontally, with different parts of the model running on the same data. In this way, Data + MLOps techniques increase operational efficiency for H3 technology providers.

It is now believed that companies like OpenAI and Google are harnessing generative AI methods to make their MLOps pipeline even more sophisticated, creating a meta-robot that can build even better robots.

For instance, ChatGPT's efficiency comes from chaining together several distinct models — starting with a regular large language model, creating a reward model with human feedback, and finally using reinforcement learning with

Figure 6. PolyAI services portfolio at Infosys

Managed models, datasets, pipelines, and end points	Frameworks (orchestration, technologies, etc.)
Open models	Open frameworks, including LangChain, Haystack, Ray Serve etc.
Closed models	Hyperscaler frameworks, including SageMaker, Azure OpenAI etc.
Hyperscaler models	Closed models (APIs)

Source: Infosys

human feedback (RLHF). This removes the operational burden from their MLOps teams as there is no need for the data engineering tasks used in H2.

In H3, the big question is whether MLOps will become obsolete for firms using ML technology. Will it fade into obscurity, or will it evolve to suit the needs of users of LLMs and generative models?

We believe that, even in organizations that buy out-of-the-box generative AI solutions, H3 models will be tailored to specific use cases, and will require MLOps to bring all components together to reduce operational complexity and increase velocity of AI products. Systems will be created that integrate several generative AI models, forging a fusion of models that's greater than the sum of its parts.

With this in mind, firms should implement maximum automation across the entire gamut of data engineering and model life cycle management, ranging from training and inferencing to API abstraction and toolkit engineering.

This is vital due to the upcoming LLM landscape that includes content such as text, images, audio, etc., and multiple models that operate on disparate data.

To make this data mesh operational, we require more advanced AI factory operations that leverage MLOps. AI engineering life cycle management, part of our PolyAI suite of services (Figure 6), which sits under Infosys Topaz, is an Infosys approach to MLOps that enables data scientists to use ML development tools of their choice and train and deploy their models at enterprise scale without having to deal with engineering complexity.

The approach also supports multiple versions of leading AI frameworks such as TensorFlow, PyTorch and others, and maintains traceability of model artifacts while at the same time enabling versioning and sharing of artifacts among development teams.

AI systems that evolve

AI systems should also be built to evolve, or improve, with time. In [Data + AI Radar](#),¹⁹ we introduced the **SURE** taxonomy. Here, an AI system moves from **S**ensing, to **U**nderstanding, to **R**esponding, and finally to **E**volving. Evolve, therefore, is the most advanced type of AI system, with models that are self-supervised and incorporate RLHF.

At present, only 15% of firms (Figure 7), including the cloud giants and others such as Apple, Meta, OpenAI, and Netflix, are able to achieve these top-level evolutionary design capabilities.

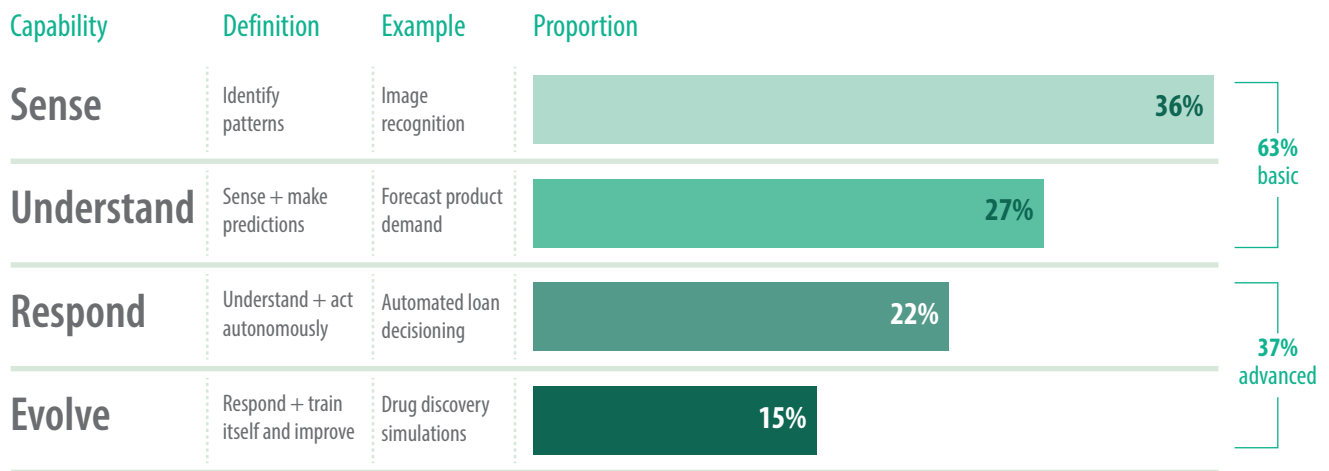
However, according to Sunil Senan, Infosys senior vice-president and head of data and analytics, "Companies need advanced AI if they are to achieve the loftiest ambitions of AI and stand out from competitors."

So what is an "Evolve" system, one that can respond, train itself, and improve?

For enterprises to leverage the foundation models used in generative AI, they need to do three things: first, they must acquire up-to-date knowledge; second, they need to perform advanced reasoning; and third, they must use actuation to make them more useful, such as automating business workflows.

1. If the system doesn't have enough knowledge, this knowledge will have to be continually updated by the system from the outside-in. This could be enterprise domain knowledge or data gleaned from searching the internet. OpenAI stopped training GPT-4 in September 2021, so ChatGPT, which is based on that large language model, has no knowledge of events after that date. To hedge against poor outcomes, OpenAI uses evolutionary design principles: ChatGPT 4 has a modular plugin architecture so that other applications can plug into it and provide additional services, including up-to-date knowledge or insights.

Figure 7. Only 15% of firms achieve evolutionary AI design capabilities



Source: Infosys

- If the question requires advanced reasoning capabilities, then the evolutionary architecture can use chain-of-thought (CoT) prompting — a series of intermediate reasoning steps — that increases the ability of LLMs to perform complex reasoning.
- Plugins are also available. If you want to book a flight, Expedia can plug in to OpenAI's architecture.

These systems must be trustworthy. For this, we need external control. An external control system — sort of like a master control plane such as [LangChain](#)²⁰ — feeds the question or prompt to the foundation model and uses APIs to orchestrate behind-the-scenes plugins, gets the answer, and feeds the response back to the user.

Further, evolutionary systems such as OpenAI also use human preference data (such as asking for thumbs up/thumbs down prompts) to continuously revise itself and offer answers that get better over time. This is an example of [RLHF](#),²¹ which drives improvement in the model over time.



Companies need advanced AI if they are to achieve the loftiest ambitions of AI and stand out from competitors.

Sunil Senan

Senior vice-president and head of data and analytics, Infosys



This is one kind of evolutionary design, using a modular architecture to iron out deficiencies. Another approach, used by Microsoft's Bing and DeepMind's Sparrow, is to continuously update knowledge by retrieving metadata. Bing does periodic searches daily — retrieving information and then sending it to the LLM for training or fine tuning.

Building block 3: Responsible AI by design

AI is now part of our lives, from building products for businesses and consumers to scanning resumes for recruitment and managing remote workers, and from drug discovery to diagnostics. Yet security and governance are only just catching up with the explosion of AI functions.

We are now in the third wave of AI evolution (H3). The first wave (H1) was driven by machine learning, and the second (H2) by deep learning. This third wave is led by foundation models trained on broad data that use self-supervision at scale and which can be adapted to perform a wide range of tasks. In short, these are the models that underpin generative AI, including the large language models that drive tools such as Google's Bard and OpenAI's ChatGPT.

This third wave of AI evolution has brought a new range of concerns to already well-articulated worries about transparency, explainability, human oversight, compliance, and continuous improvement. Generative AI creates copyright concerns around the billions of parameters used to train large models, as well as fears about perpetuating bias and disadvantage, malicious use of AI-generated content, and limited access to foundation models and training data or

weights. This limits our ability to address the underlying limitations of these models. We neither choose the data these models are trained on, nor provide human preference data used in RLHF.

What we can control is a careful evaluation of the model outputs to actively search for biases or mistakes.

Appropriate governance

Therefore, appropriate governance is a key plank to becoming an AI-first organization.

At the heart of ethical AI is the concept of “responsible by design.” This is already familiar to cybersecurity professionals, who adopt this framework to create and deploy security products and policies. The aim is to bake in security — and now AI ethics — at every step of the process.

The principles to guide responsible by design AI include:

- Human oversight and governance at every stage.
- Constant auditing of processes and products for fairness, inclusiveness, and prevention of harm.

- Transparent and explainable AI that is reliable, safe, secure, and private at each stage.

How should organizations implement these principles?

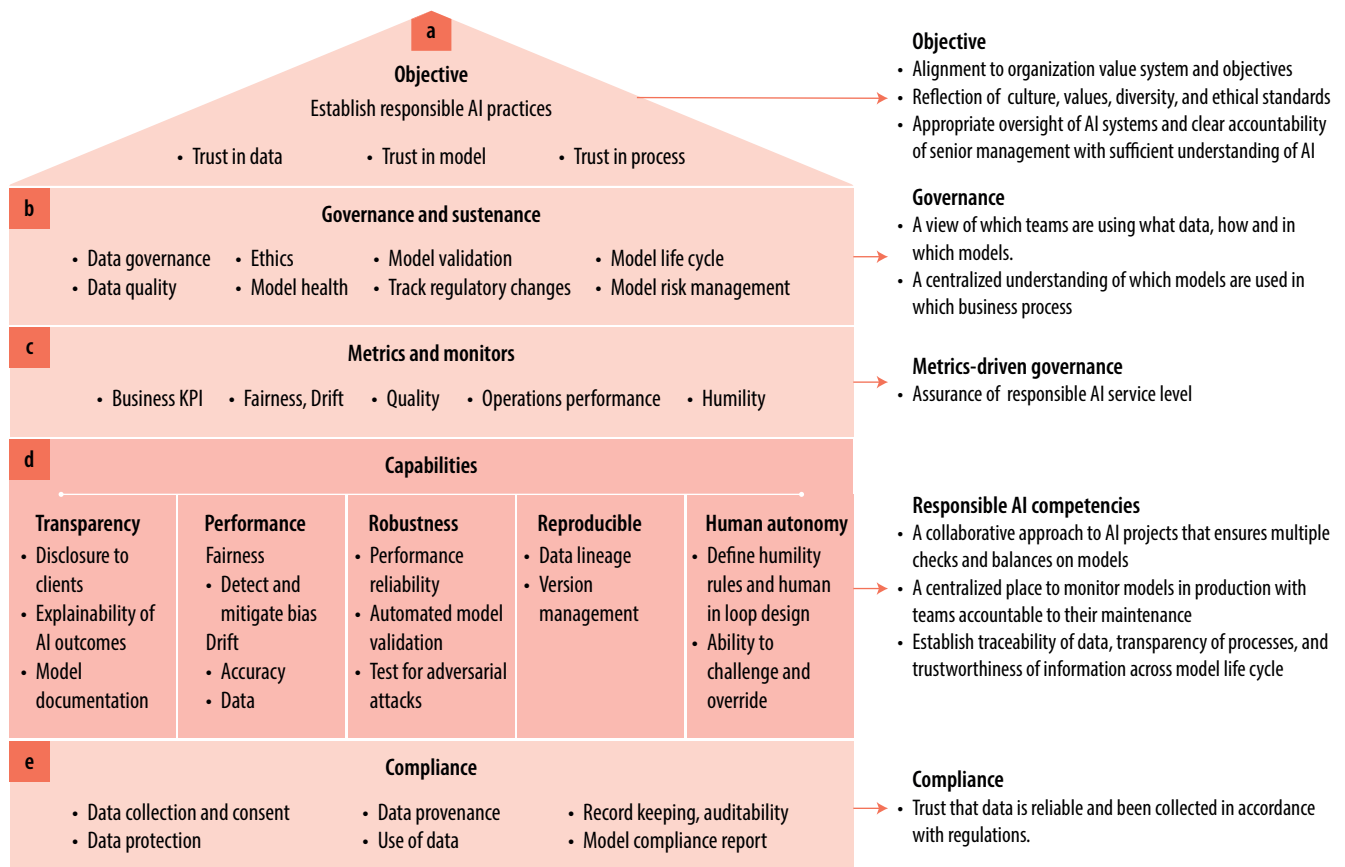
We have identified five responsible-by-design building blocks: objective, governance, metrics, capabilities, and compliance (Figure 8).

Included in these building blocks are a focus on aligning with the organization’s wider objectives, value systems and diversity standards, as well as details of governance such as assuring data quality, validating the model and tracking regulatory changes.

For human oversight and governance, organizations should identify a diverse range of stakeholders at the outset and engage with them regularly.

Having a wide range of voices that is heard and acted upon will raise potential issues early, and their warnings should mitigate any potential harm threatened by use of AI products or services.

Figure 8. Five responsible-by-design building blocks



Source: Infosys

It's vital for organizations to assign responsibility for human oversight, making sure that business sponsors from the outset understand the intended purpose of the AI. This includes compliance teams tasked with making sure that the AI processes and products meet regulatory requirements and do not perpetuate biases. It also includes Ops and IT teams making sure a model can be explained to regulators and is continuously monitored for accuracy.

Continuous auditing

Auditing requires a cross-functional approach: it cannot be left to solely to product and tech teams. It means working with legal teams, with data protection and with cybersecurity teams to review implementation across the business and to consider how AI products and processes must comply with laws and regulations.

Continuous auditing also means continuous feedback so that engineering teams respond to problems that arise, such as bias in AI decision-making and hallucination in generative AI chatbots. A moderation function can continually review the AI's output and flag problems for fixing.

Transparency means that the algorithms and inputs that led to a decision can be checked and understood by all stakeholders of the AI product

For foundation models and large language models in Horizon 3, an external control system can also mitigate unintended hallucinations. Currently a few external control platforms facilitate flow between these large language models alongside external sources to augment the model's knowledge, reasoning, and actuation.

External controls also create the necessary safeguards for responsible AI design. This is why prompting (see the AI operating model section) is important to promote responsible AI design and development practices.

AI products and processes must be transparent and their outputs explainable. This means that the algorithms and inputs that led to a decision or other output can be checked and understood by humans in the business, and by customers and users outside the business.

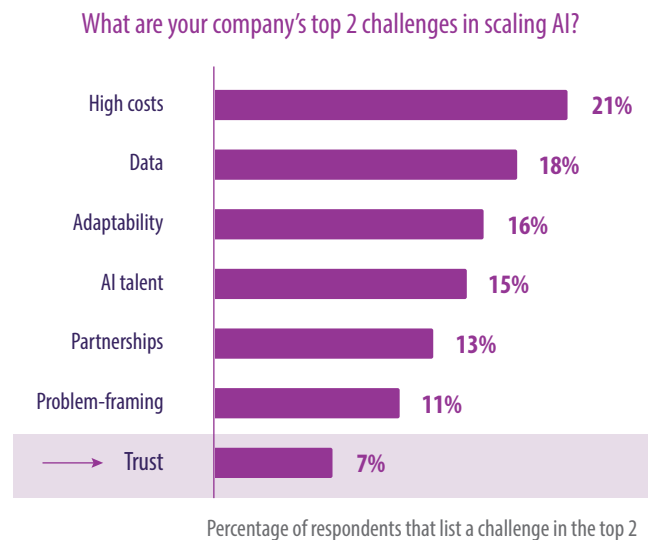
This increases trust in AI systems, which is a foundational prerequisite for deploying advanced AI systems.

Trust: the next big challenge

As we wrote in [Data + AI Radar](#),²² trust is the next big challenge in implementing AI systems. When AI systems are deployed at scale, "trust and responsible AI systems become a major issue," says Bonnie Holub, a data science leader with Infosys Consulting. "We see trust [and responsible AI] as crucial parts of the non-financial governance issues investors are demanding from companies," she adds.

However, despite its importance, executives rated trust as a low concern when surveyed for Data + AI Radar (Figure 9).

Figure 9. Despite its importance, executives rated trust as the lowest concern



Source: Infosys

“

Trust and responsible AI are crucial parts of the non-financial governance issues investors are demanding from companies.

Bonnie Holub

Data science leader, Infosys Consulting

”

Reliability, safety, security, and privacy occur when AI tools, products, services, and processes are built to robust standards and operate consistently in the way they were originally designed.

These systems should also continue to work as designed under unexpected conditions, and regular testing for reliability must be a part of the design, implementation, and maintenance processes.

An organization can only consider itself an AI-first entity if it has embedded responsible by design principles into every corner of its business and work. This applies to those building these tools and those using and deploying them.

Only then can an organization be truly ready to be a part of this transformative and exciting landscape.

Building block 4: The AI operating model

AI-first talent

AI talent is one of the top four challenges for executives to transform their enterprise to AI-first (Figure 9). What specific talents and skillsets should they seek?

Prompt engineering is a key skill required. Anthropic, Google's latest \$300 million investment, is hiring prompt engineers at salaries up to \$335,000 a year, underscoring the value placed on prompt engineering and its role as a crucial skillset.

Prompt engineering is not easy. Orchestrating systems is a complex, multifaceted task that requires a deep understanding of the interplay between foundational models, external systems, data pipelines, and user workflows.

Workers in this new normal will discover, test, and document best practices for a wide range of tasks customers use when collaborating with AI.

Prompt engineers will also build a library of high-quality prompts or prompt chains and build tutorials and interactive tools to pass that knowledge to others. According to experts at Infosys, firms should seek people with a hacker spirit who love solving puzzles, have excellent communication skills to teach both human and machine technical concepts, and who are familiar with the architecture and operation of large language models.

In our previous work on AI, [Tech Compass](#),²³ we examined the skills required for the AI horizons: H1, H2 and H3.

Previously, these roles required data scientists who majored in mathematics and econometrics for H1, while the requirement shifted toward data engineers for H2. Now, with H3 in sight, organizations need programmers who can craft prompts to elicit trustworthy responses from foundation models and interact with external third-party systems. For instance, the prompt "Buy me an iPhone and send it to X address" will only be effective if a robust orchestrating framework exists, allowing the large language model to seamlessly integrate with an e-commerce giant like Amazon.

With much of the AI power vested in a few companies such as Meta, Google, OpenAI, and Microsoft, the need to learn prompt engineering will ripple out to the whole AI-first organization. It's important to recognize that some of the existing jobs will get displaced and new roles like prompt engineers and model tuners will be created. Existing roles will have to be enhanced to use AI tooling and twins to make themselves hyper-productive, in addition to softer skills and characteristics — including empathy, creativity, problem solving, and integrity.

For now, it's beneficial to think about the AI-first skills dimension through three prisms. The first is value: people who can scale AI throughout the organization will be in high demand, as there is little value to use AI in pockets.

Second, value must be buttressed by trust, as we discuss in our [Data + AI Radar](#).²⁴ Firms should employ experts who understand the privacy, security, and legal aspects of AI.

Third, for those systems not yet in H3, data engineers with cross-domain knowledge are required, as well as AI architects with both depth and breadth of software engineering knowledge. This includes understanding containers and code dependencies, peer code review practices, coding standards, logging techniques, clean code and code optimization, and code modularity. At the same time, existing ETL approaches taken by software engineering must be enlarged to support AI trust, risk, and security functions. Organizations should look at areas such as bias, differential privacy, data quality and capabilities around synthetic data generation.

In all of this, firms must understand that AI-first is a mindset shift that everyone in the business needs to embrace. AI is even more fundamental than mobile, and cloud, and good talent will realize that they have to take advantage of AI.

"While some worry that AI will take their jobs, someone who is an expert at AI will certainly do so," Jensen Huang, CEO of NVIDIA, the GPU company, recently told Infosys. "AI will supercharge the performance of programmers, designers, artists, marketers, and manufacturing planners," he added.

Become product-centric to succeed

AI-first organizations must also be increasingly product-centric. They will organize the firm around dedicated customer journeys or value streams, rather than traditional functions, so AI products are delivered at the pace required (Figure 10).

Product-centricity is less about products and more about the **value delivered**.²⁵ AI products are owned by product managers, a role already in demand. Our **Agile Radar research**²⁶ found that 74% of C-suite and IT executives across the US and Europe invest in product management, underlining it as a key business priority.



AI will supercharge the performance of programmers, designers, artists, marketers, and manufacturing planners.

Jensen Huang
CEO of NVIDIA



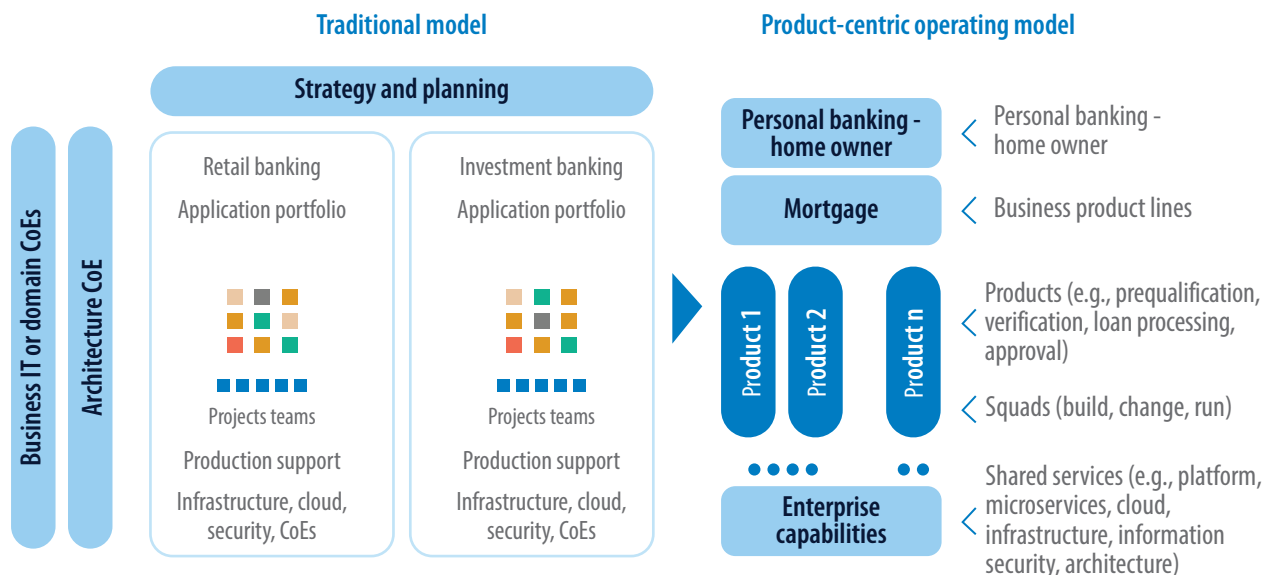
There are many benefits to the product-centric approach. AI-first, product-based, data-driven firms take advantage of business opportunities by tracking what works well in real time and feed the insights back into product design.

AI projects are iterative and based on continuous learning, as demonstrated in this report's section on engineering excellence. In the product-centric operating model, AI product teams are long-lived and fully invested in growth beyond simply launching the capability.

Data and AI product teams integrate pipelines and roles and responsibilities are defined to ensure handoffs are seamless and schedules are aligned. At Peloton, for example, new data captured on AI-based virtual fitness studios is used to build on product development. AI teams thus develop an ongoing understanding of customers and users, which in turn leads to better design research processes. This human-centered approach then spurs more innovation, driving deeper, richer end-user engagement and business outcomes from AI.

AI-first, product-based, data-driven firms take advantage of business opportunities by tracking what works well in real time and feed the insights back into product design

Figure 10. In the product-centric operating model, IT delivery is aligned to customer journeys



Source: Infosys

This operating model ensures that IT is a stronger force in the business, as it becomes integral to product strategy and that it is vested in the success of AI.

As we discuss in our [recent paper on product-centricity](#),²⁷ this also creates new business models and revenue sources, and paves the way for a platform ecosystem to flourish.

For example, OpenAI and Google have ensured that their AI chatbots also become platforms for other products that are easily integrated, generating significant additional data and value for their businesses.

Scaled engineering

Product-centricity, however, doesn't detail how core AI engineering, tooling and playbooks are used. For this, the concept of "hub-and-spoke" is useful and is analogous to the platform engineering methodology introduced earlier in this report.

As a middle ground between centralized and decentralized AI and data usage, a hub-and-spoke approach can provide the agility and consistency that AI product-based teams need.

This approach introduces a central team or center of excellence (the "hub") that owns the AI and data platform, tooling, and process standards. This is complemented by business teams (the "spokes") that own the AI products for their domains.

This approach resolves the "anything goes" phenomenon of decentralized AI team topologies while empowering subject matter experts (SMEs), or AI stewards, to independently create AI products that cater to their specific needs.

For this approach to be scalable, a common platform that supports AI model sharing, with conformed dimensions, collaboration, and ownership, is critical, and is similar to platform engineering.

Common models and dimensions, such as time, product, and customer, are established, while the domain experts own and define their business process models. This enables self-service and model re-use, increasing efficiencies and innovation by allowing product owners and domain experts to combine their AI models with models from other domains to create new mashups for answering deeper questions.

Further questions to consider include: How do data and analytics teams work with software engineering and AI teams on the data pipeline? And how do firms accelerate the volume of feedback that they can incorporate from domain experts?

Microchange management

As AI booms, executives might struggle with the change management that's required. From rethinking the operating model to establishing governance; from upskilling teams so that they understand responsible AI practices, to performing core engineering work, the process of assimilating AI into an organization involves many complex and interdependent moving parts. Few executives are able to measure product KPIs from pilot to launch and to guarantee a seamless integration of AI.

Microchange management (or "micro is the new mega"^{28m}) is a way to overcome this inertia. Instead of making drastic changes all at once in a big-bang program, cross-functional product-based teams deconstruct work into a series of small components. These teams iterate change and achieve adoption through an Agile cadence.

OpenAI and Google have ensured that their AI chatbots also become platforms for other products that are easily integrated, generating additional value for their businesses

AI-first organizations use microchange for core engineering and applied AI, and to change employee behavior (upskilling, for example) through slight modifications to habits and routines.

This is important where product-based culture needs to catch up with advances in AI-first engineering. AI products are piloted on a small scale, and then lessons from that pilot scheme can then be used to refine and then scale the roll-out, whether for a generative AI chatbot or IDP.

According to Infosys research, realizing the full potential of change management demands a detailed, meticulous approach that prioritizes high-value use cases, that starts with small-scale implementations, and which emphasizes the importance of people and processes over the technology.

Here, microchange management provides a low-risk approach to turning complex transformations into manageable, bite-sized changes, thus "minimizing the leap of faith required to reach the other side," as authors Jeff Kavanaugh, head of Infosys Knowledge Institute, and Rafeef Tarafdar, Infosys CTO, [write](#)²⁹ in the Harvard Business Review.

With time, this leads to real AI adoption, the overarching goal of an AI-first organization.

Infosys Topaz: Accelerating AI-first business value for Infosys and our clients

Infosys
topaz

As this Tech Navigator discusses, now is the time to go AI-first. But it is a fundamental change to how businesses should organize themselves and work. At Infosys, we make sure we thoroughly understand the transformation that's needed. In our Live Enterprise transformation, we did everything internally before working with clients. But with our AI-first push, powered by [Infosys Topaz](#), we are doing everything in parallel, using Infosys as a sandbox, so that clients can benefit from our experience as well as the full potential of the technology. In this way, we will help our clients move from cloud-led to ecosystem- and AI-led, accelerating growth.

First, important in all this is that we don't see AI-first as a replacement of people. Rather, we want to use this transformative tech to amplify human potential. An example of this is how we use AI assistants that are trained on the enterprise knowledgebase and body of work across the organization. This unlocks value, allowing humans to do work productively, creating impact for the whole organization.

Also important, and we've dedicated full section of the report to this, is the idea of responsible by design. This covers privacy, security, explainability, traceability, bias detection, IP and regulatory compliance. A lot of work we're doing this year is to make sure the AI products we build are responsible by design.

As we take our AI-first strategy, through Infosys Topaz, to our clients, a few things will stand out:

- Every person in the firm will get an AI assistant. If you're a developer, you will get a code assistant to use to generate and complete code, write test cases etc. If you're in support, you will use the AI assistant to identify root causes and understand parts of your job you may not be aware of.
- PolyAI, as we've described, will allow a lot of this new technology to evolve – it means we as a firm can operate and pick the right AI provider and the right model to perform any operation.

- Any activity will have "AI in the flow". This means that employees don't have to use separate screens and plug-ins all the time – the AI just helps them in their work, right where they are.
- Building AI first talent is key to success, and enablement and training is required across three levels – awareness, AI solution building and deep tech AI engineering.

We have taken the strategy, architecture, design and learnings, and best practices from our AI-first implementation, coupled with the insights and learnings from client projects, to bring together Infosys Topaz.

Infosys Topaz brings in knowledge assets, specialized IP and platforms (150-plus pre-trained AI models and more than 10 AI platforms), AI talent (AI-first specialists and data strategists) and a partner and innovation ecosystem. In this way, Topaz helps customers navigate from a digital core to an AI-first core, providing them with cognitive capabilities, and helping them capitalize on generative AI opportunities.

As an example of some of the work Infosys Topaz has accomplished with clients:

- We've created a personalized recipe generator for dietitians who are making recommendations based on ingredients purchased, age, diet type the user is on, and geography.
- We created an efficient process for one manufacturing firm to search and map data on parts, and display on their site for easy reference using the text processing capabilities of GPT-3.
- We have created an advisory service to help wealth managers make the right offerings to customers and be more relevant to their needs.

At Infosys, we are going AI-first so that we understand how to help your business go AI-first. To find out more, infosystopaz@infosys.com

References

1. Tech Navigator 2022: Building the human-centric future, Harry Keir Hughes, 2022, Infosys Knowledge Institute.
2. On the opportunities and risks of foundation models, Rishi Bommasani et al, July 12, 2022, Stanford Institute for Human-Centered Artificial Intelligence (HAI).
3. Advanced trends in AI: The Infosys way, Sudhanshu Hate, Harry Keir Hughes, Jeff Mosier, and Anu Mary Tom, October 1, 2020, Infosys Knowledge Institute.
4. Infosys Digital Radar 2023: The next digital frontier, Harry Keir Hughes, 2023, Infosys Knowledge Institute.
5. Artificial Intelligence index report 2023, Nestor Maslej et al, 2023, Stanford Institute for Human-Centered Artificial Intelligence (HAI).
6. The state of generative AI in 7 charts, January 25, 2023, CB Insights.
7. Data + AI Radar 2022: Making AI real, Chad Watt, 2022, Infosys Knowledge Institute.
8. Tech Navigator 2022: Building the human-centric future.
9. Infosys Digital Radar 2023.
10. Generative AI landscape: Potential future trends, George Lawton, April 19, 2023, TechTarget.
11. Microsoft invests \$10 billion in ChatGPT maker OpenAI, Dina Bass, January 23, 2023, Bloomberg.
12. Infosys Tech Compass, 2022, Infosys Knowledge Institute.
13. Platform engineering, Vishwanath Taware et al, 2023, Infosys.
14. What is platform engineering?, Lori Perri, October 5, 2022, Gartner.
15. The Live Enterprise: Create a continuously evolving and learning organization, Jeff Kavanaugh and Rafee Tarafdar, 2020, Infosys.
16. A self-driving sustainable cloud for greener business, Rajeshwari Ganesan, Professor Ravishankar K. Iyer, and Harry Keir Hughes, March 2, 2023, Infosys Knowledge Institute.
17. Infosys Digital Radar 2023.
18. Scaling AI: Data over models, Rajeshwari Ganesan, Sivan Veera, and Harry Keir Hughes, May 3, 2021, Infosys Knowledge Institute.
19. Data + AI Radar 2022.
20. Welcome to LangChain, Harrison Chase, 2023, Python.
21. Illustrating reinforcement learning from human feedback (RLHF), Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla, Dec. 9, 2022, Hugging Face.
22. Data + AI Radar 2022.
23. Infosys Tech Compass.
24. Data + AI Radar 2022.
25. Product-centric value delivery: A new digital strategy, Harry Keir Hughes, Parag Palshikar, Madan Gopal Malladi, and Ayyapa Das, February 21, 2022, Infosys Knowledge Institute.
26. Agile Radar 2021, Harry Keir Hughes, 2021, Infosys Knowledge Institute.
27. Product-centric value delivery: A new digital strategy.
28. Break down change management into small steps, Jeff Kavanaugh and Rafee Tarafdar, May 3, 2021, Harvard Business Review.
29. Break down change management into small steps.

Contributors

Authors

Rajeshwari Ganesan | Distinguished Technologist, Infosys

Rajeev Nayar | CTO of Data & AI, Infosys

Kamalkumar Rathinasamy | Distinguished Technologist, Infosys

Rafee Tarafdar | CTO, Infosys

Kate Bevan | Infosys Knowledge Institute

Harry Keir Hughes | Infosys Knowledge Institute

Acknowledgements

Syed Ahmed

Balakrishna DR

Uday Kumar Gupta

Karthik Andhiyur Nagarajan

Sundaresan Poovalingam

Priya Pravas

Pragya Rai

Shreshta Shyamsundar

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



© 2023 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and / or any named intellectual property rights holders under this document.

