BANK OF AMERICA INSTITUTE

BANK OF AMERICA

Transformation

Next Gen Tech: Artificial intelligence

08 May 2024

Key takeaways

- While generative artificial intelligence (AI) and large language models stole the headlines in 2023, the technological landscape will shift focus to the implementation of AI tools and applications, with five breakthrough AI technologies leading the way.
- End-device AI, enriched simulation, knowledge graphs, hyperdimensional computing, and artificial general intelligence could enable an abundance of opportunities beyond the digital to the physical domains of end-devices, robotics and life sciences.
- Bank of America Institute's 'Next Gen Tech' series explores 30 breakthrough technologies across AI, computing, robots, communication, healthcare, energy and mobility, that are about to alter our lives. Join us as we discuss what's next on the tech horizon.

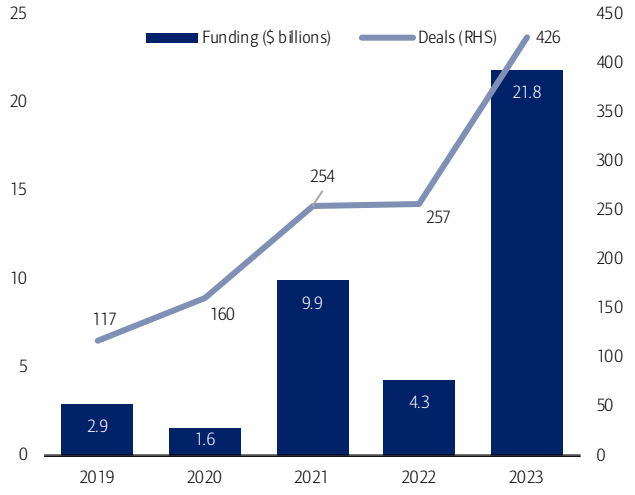
This publication is part of Bank of America Institute's 'Next Gen Tech' series – focused on sharing 30 breakthrough technologies that will transform the world. Each week, we'll highlight one of seven categories (artificial intelligence, computing, robots, communication, healthcare, energy and transport), and share advancements within each, so stay tuned for more.

AI for all

In our series introduction, Next Gen Tech: Breakthroughs that will transform the world, Bank of America Institute discussed how rapid shifts in innovation are transforming businesses and the world. In fact, we noted that the fastest transformation in human history is ahead of us. Here, we discuss the first of seven categories of Next Gen Tech breakthroughs – artificial intelligence (AI).

The pervasiveness of AI is at an 'iPhone defining moment.' Last year, generative AI (genAI) stole the headlines – but BofA Global Research underscores that this is just the beginning and this 'AI revolution' will only accelerate from here. The buzz all started with the release of ChatGPT by OpenAI in November 2022, and in turn, 2023 saw a surge in genAI investment (Exhibit 1).

Exhibit 1: Disclosed equity funding (US$ bn) and number of deals
Generative AI investment surged in 2023



Source: CB Insights: State of AI 2023 Report, BofA Global Research

BANK OF AMERICA INSTITUTE

Since ChatGPT's release, a variety of both closed- and open-source models have been introduced, with companies already starting to develop, adopt or integrate AI into their products or businesses. The momentum will only intensify from here, with

more AI tools (e.g., simulation, knowledge graphs) and applications likely to be available soon, which could enable an abundance of opportunities beyond the digital to the physical domains of end-devices, robotics and life sciences. So, what comes next in the AI space?

1) End-device AI

Fuelled by ChatGPT's development, cloud AI was a key topic over the last year. However, on top of cloud AI opportunities, end-device AI is growing in importance because AI functions are necessary on both the end-device and cloud sides to make the most of AI. With a significant amount of data generated from numerous devices every day, the key consideration is where the data should be processed. Multiple industry suppliers define three categories within the compute spectrum:

- **End-device AI** refers to the deployment of AI functions/models on local devices, e.g., smartphones, autos, and wearables. Due to the long data transmission path, transferring data from end-devices to the cloud introduces costs and other issues including longer latency, power consumption for data transmission, bandwidth, server capacity, private information leakage, etc., which can lower the service quality. Therefore, end-device AI helps share the power load of large servers to improve the performance of the broader AI ecosystem.

  In this publication, the definition of end-device AI computing is: collecting data from numerous end-devices like smartphones, cars, white goods (large appliances), security cameras, wearables, and streetlights, etc., and then processing the acquired data on end-devices (with embedded AI chips/cores) or sending that data to a gateway or another processing device near them.

- **Edge AI:** Based on the abovementioned definition, edge AI offloads AI and machine-learning (ML) processing from the cloud to powerful servers at the edge of the network, such as offices, 5G base stations, and other physical locations near the connected endpoint devices.

- **Cloud AI:** The collected data/prompts are sent back to data centers, and then the end-devices grab the calculated results from big data pools on the cloud.

Benefits of end-device AI

Exhibit 2: Comparison of end-device AI, edge AI, and cloud AI
End-device AI's advantages include lower latency, higher power efficiency, limited cost, and higher context awareness

Features	End-device AI	Edge AI	Cloud AI
Use cases	Smartphones, vehicles, white goods, security cameras, wearables, streetlights, etc.	On-premise servers, base stations, IoT (internet of things) gateways	Data centers
Latency	Low	Medium (10us-10ms+)	High (100ms+)
Bandwidth required	Low	Medium	High
Processing power	Low	Medium	High
Storage capacity	Low	Medium	High
Security	High	Medium	Medium (cloud back-up)
Computing cost	Low	Medium	High
Context awareness	High	Medium	Low
Power efficiency	High	Low	Low
Maintenance & Upgradeability	Medium	Medium	High (centralization)

Source: BofA Global Research

BANK OF AMERICA INSTITUTE

The key features of end-device AI compared with edge/cloud AI include lower latency, higher power efficiency, limited cost, and higher context awareness (Exhibit 2), which should translate into the following benefits, according to BofA Global Research:

- **Faster response time**: Some simple AI tasks only require AI computing with local storage capacity on the end-device side. In such use cases, adopting end-device AI could lead to faster response times when compared to capturing data from the cloud.

- **Better accessibility for consumers**: Given end-devices are physical and closest to end-users, the AI processing on them could lead to better accessibility.

- **Better privacy and communication security**: With higher security requirements on the end-device side (e.g., face recognition/fingerprint identification, one-time password, text password, etc.) for personal information protection, the security level could be higher if data remains local.

- **Offloading from cloud**: Edge devices with computing power could share the cloud's burden to improve the overall AI service.

More achievable due to wireless connectivity (e.g., 5G, Wi-Fi) & IoT (internet of things)

Three technology megatrends – wireless communication, AI, and IoT – are mutually driving the growth. 5G offers high speed, low latency, and wide range connection for AI, especially on the end-device side, while AI computing could enhance 5G's transmission performance and efficiency.

Challenges for end-device AI

To perform end-device AI tasks locally and connect with the edge and/or cloud for more complex AI tasks, there may be some key challenges:

- **Power consumption**: End-devices like smartphones, smart watches, etc. are close to end-consumers and running on battery power, and thus power consumption needs to stay at low levels even if equipped with more features.

- **Cost**: The increasing semiconductor content needed to support more AI functions should inevitably lift the cost of end-devices.

- **Algorithm/software**: As end-devices have limited resources, including processing power, memory, and storage, vs. the cloud or a data center, the AI/ML algorithm and related software must be optimized to work within these constraints.

- **Security**: One of the most significant challenges of end-device/edge AI is data privacy disclosure. End-devices and the edge servers store and process a large amount of data, including sensitive personal data, which makes them attractive targets for cyberattacks.

2) Enriched simulation:

Enriched simulation uses AI to accelerate the discovery process and identify the most viable simulations, speed up the creation of new molecules and bring down the cost to do so.

Many of our everyday products are complex and over time designers have come to rely on computer-driven simulations but they often take time to run. Even once possibilities are found, additional simulations need to be run to ensure safety. AI simulation combines techniques from quantum physics and deep learning to enable sampling a vast dataset quickly and efficiently. AI and simulation technologies bring the ability to take a molecular structure and simulate it billions of times, making small changes each time to see which structure is optimal. We can now do this in a matter of weeks and months – a task that would take 10 years in the physical world.

What does the simulation process involve and how does it work?

Bayesian statistics create methods that rival traditional simulation methods by finding the optimal solution based on the limited information that the system has. In this way, the method minimizes the uncertainty of the next best solution thing using existing knowledge to select the next best parameters.

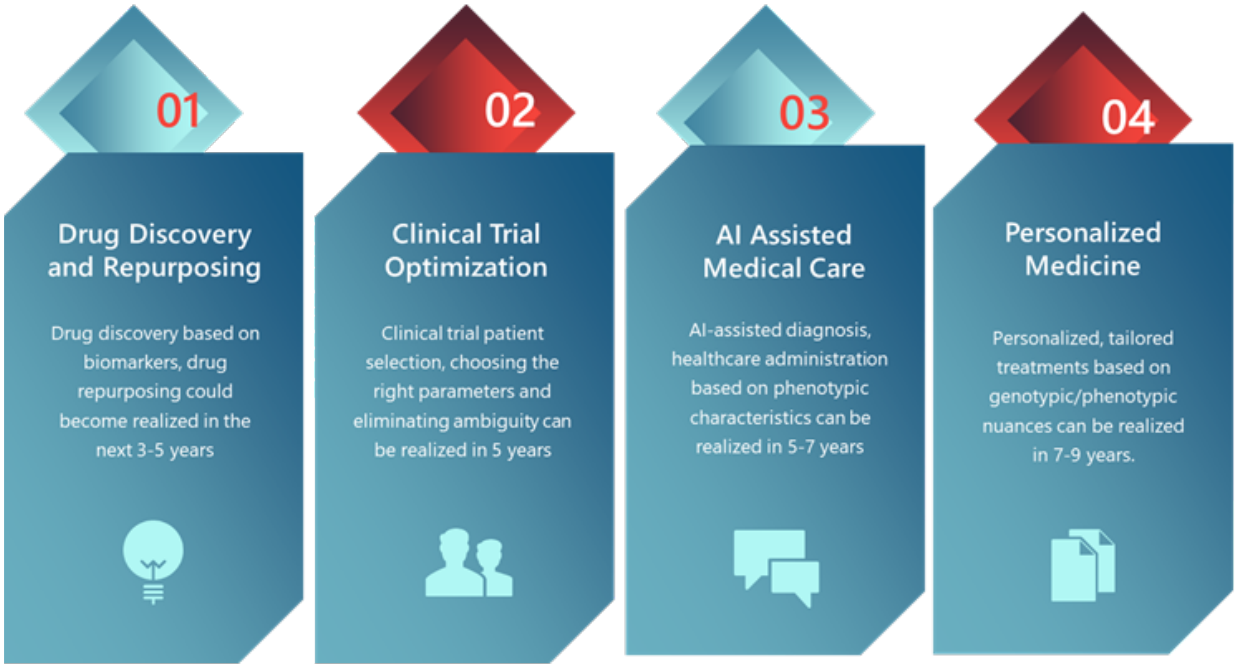
Let's talk applications

There are numerous applications for enriched simulation including drug discovery, chip design, advancing the discovery and design of innovative chemicals and materials, and enabling predictive analytics and streamlining operations (risk management tools, in particular) within the finance industry, to name a few.

- **Drug discovery** involves high costs and has a high failure rate. In fact, the average investment is $1-4 billion and typically takes 10-15 years to develop new drugs. Additionally, there is a 90% failure rate, which makes the business model unpredictable. In turn, revenues from the 10% of drugs that are successful have to pay for the 90% that failed.

  The high failure rate helps explain Eroom's law (the observation that drug discovery becomes slower and more expensive over time), which implies a decrease in R&D (research and development) efficiency. Eroom's law often results in companies spending more per new drug on R&D than they make in revenue. New AI simulation work can change sectors such as life sciences from a business of mostly failure to one with more predictable revenues, as it allows companies to take all the data and molecular information and run billions of simulations de-risking the molecules, reducing drug development time dramatically (Exhibit 3).

Exhibit 3: Timeline on Future AI Development in Healthcare
BofA Global Research sees the application of AI in drug discovery realizing in the next three to five years, while other healthcare applications will take longer at seven+ years



Source: BofA Global Research

BANK OF AMERICA INSTITUTE

- **Chip design:** Electronic design automation (EDA) vendors have made tools for chip design using rule-based systems and physics simulation. But now, AI can help chipmakers push the boundaries of Moore's Law (the observation that the number of chips in a circuit doubles every two years) further. Simulation can design chips faster than older methods and make new and improved chips. All in all, these tools can increase supply chain security and help mitigate shortages.
- **Chemicals and materials:** Simulation can also be used to advance the discovery and design of innovative chemicals and materials. Computing power now allows us to take the molecular structure, make its digital twin and run billions of combinations to create products faster. Take the example of lithium and the ongoing supply chain issues of scaling EV battery production: what if we could find a new battery chemistry that rivals the existing types of battery chemistries? However, given the number of elements out there, we can run a simulation of say 19 elements and the 10^117 combinations that are possible to create batteries without lithium, for example. This can be simulated by a computer using GPUs to narrow down the mix of chemicals that would work and shorten development lead times as a result.
- **Finance:** The finance industry can also leverage AI and simulation by enabling predictive analytics and streamlining operations, and risk management tools. Monte Carlo simulation (a probabilistic model that can include an element of uncertainty or randomness in its prediction) is the standard process when looking at risk exposure, but this method may no longer be sufficient in today's complex world. There are several challenges such as being computationally intense and time-consuming, being dependent on the accuracy of the input data and being sensitive to random number generation. Using AI and simulation, we can put in complex portfolios and find the optimal one by conducting risk analysis, specifically tail risk (which is not analyzed in Monte Carlo simulation).

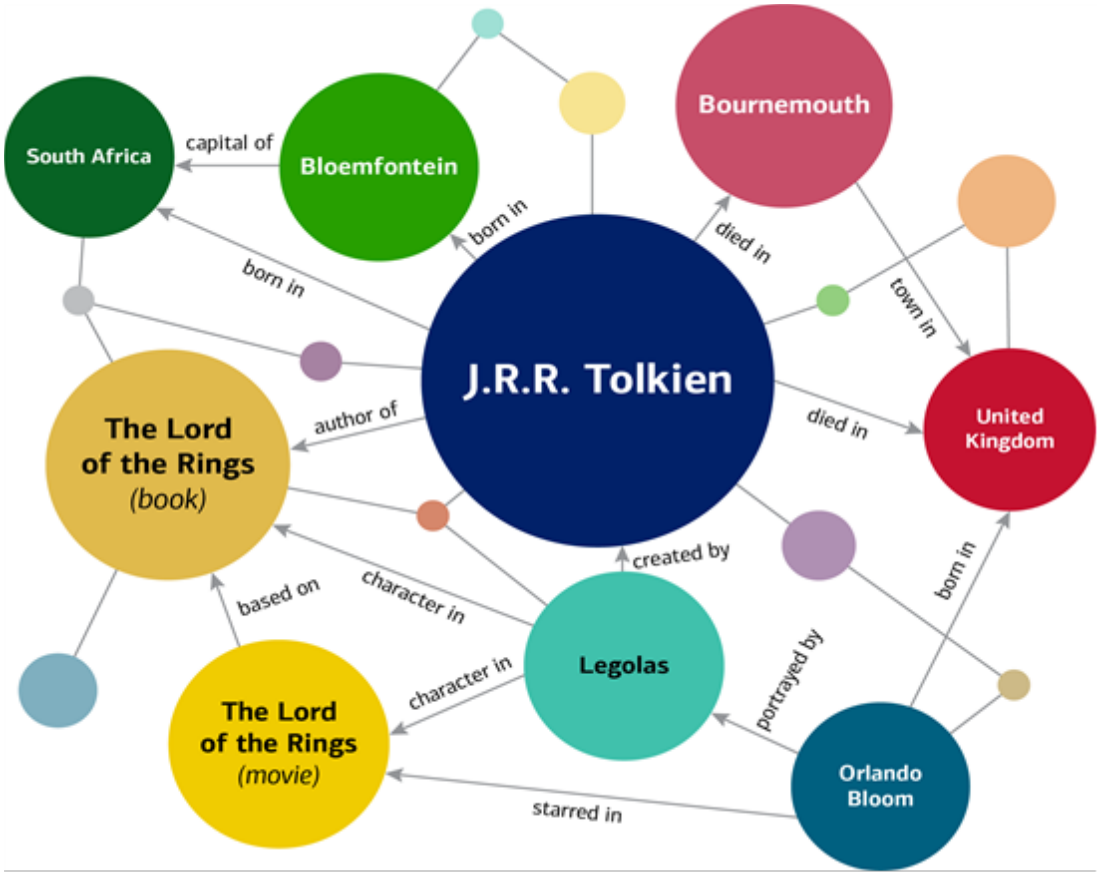
3) Knowledge Graphs:

Knowledge graphs (KGs) are a way to store information and show the relationships between related sources of information. They organize data from multiple sources, capture information about the topics and forge connections between them, as demonstrated by the example KG in Exhibit 4, below. It's important to note that not every dataset is a knowledge graph. There are many definitions, but most agree that knowledge graphs have the following characteristics:[1]

[1] Alan Turing Institute

- **Graphs:** the content is organized as a graph, where nodes (entities of interest), relationships between them and attributes of the nodes are equally important. This makes it easy to integrate new datasets and formats together by navigating from one part of the graph to another through links.
- **Semantic:** the meaning of the data is encoded for programmatic use in an ontology – the schema of the knowledge graph, which describes the types of entity in the graphs and their characteristics.
- **Alive:** KGs are flexible in terms of the data and schemas they can support. They can evolve to reflect changes in the schema and when new data is added to the graph.

Exhibit 4: Sample knowledge graph on Lord of the Rings related information, stemming from its author, J.R.R. Tolkien
Knowledge graphs organize data from multiple sources, capture information about the topics of interest and forge connections between them



Source: Semrush

BANK OF AMERICA INSTITUTE

KGs are fuelled by machine learning and use natural language processing (NLP) to construct a comprehensive view of nodes, edges and labels through a process called semantic enrichment.[2] When data goes through the system, KGs can identify individual objects and understand the relationships between them. This knowledge is integrated with other datasets, which are relevant and similar in nature.

Use cases are vast

Today, KGs are used in search engines and websites, chatbots, retail product recommenders, content platform recommendation engines, know-your-customer (KYC) initiatives, and organizing and categorizing relationships between different types of medical research withing the healthcare sector, for example.

- **Search engines and websites:** DBpedia and Wikidata are two different knowledge graphs for data on Wikipedia.org. Google's knowledge graph is represented through Google's Search Engine Results Pages (SERPs).

[2] IBM

- **Retail:** recommending products based on individual purchase behaviour and purchase trends across demographic groups
- **Entertainment:** KGs can be used for AI-based recommendation engines for content platforms based on clicks, online engagement behaviours
- **Finance:** KGs have been used in know-your-customer (KYC) and anti-money laundering initiatives within the finance industry.
- **Healthcare:** KGs can organise and categorise relationships between different types of medical research.

Why do we need KGs?

Knowledge graphs could be an important complementary technology to mitigate the problem of 'hallucination' – large language models (LLMs) providing inaccurate information with a high degree of confidence. Knowledge graphs ingest huge amounts of factual information from multiple sources forging connections between them. Integrating a knowledge graph with an LLM involves incorporating a contextual knowledge base into the model and allows it to make logical connections between concepts.

In this way, the LLM can draw on a variety of information sources, including structured and unstructured data to generate more accurate output. Knowledge graphs are not probabilistic engines like LLMs. Instead, they can help enhance LLMs by being a centralized source of accurate knowledge for inference and interpretability and they reduce the need for large, labelled datasets.

In the chemicals sector, for example, one could use KGs and LLMs to create a node for every molecular entity mentioned in literature, grouping together those that are similar, to identify patterns and opportunities. They're available mostly via open-source development tools.

In the biopharma industry, a company might want to create an LLM-based chatbot that can answer questions about clinical trials. To address hallucination, the company could combine LLM with a knowledge graph to create a detailed medical knowledge base that includes structured and unstructured information about drugs and their trials. The LLM would be able to refer to the contextual knowledge base of a knowledge graph to identify and analyse all the information related to that compound.

There are benefits to taking this approach, e.g., there is a centralized source of accurate knowledge and structured knowledge fusion of information in different formats.

4) Hyperdimensional AI Computing:

Hyperdimensional computing (HDC) uses high-dimensional vectors (lists of numbers that present information that can be combined in different ways) to represent information rather than the traditional binary system. It can capture more complex data patterns and allow computers to retain more memory, thus reducing their computing and power demands.

Artificial neural networks that power generative AI cannot reason

In 2023, we saw a large surge in interest in generative AI with uptake in popular AI platforms. However, the underlying architecture – artificial neural networks (ANNs) – has limitations, such as difficulty in reasoning. Humans can reason by analogy. When we see something new, we can infer new concepts from existing knowledge without needing to form new neurons.

Artificial neural networks require more artificial nodes to scale up statistical abilities, which allows them to learn new concepts (statistical AI). There is another competing approach called symbolic AI, which uses logic-based programming and symbols to represent concepts and rules. The challenge is to combine them to get the best of both worlds.

Hyperdimensional computing 101

HDC leverages statistical AI (draws upon statistical and mathematical foundations) whilst emulating symbolic AI (uses logic-based programming and symbols to represent concepts and rules). It is a relatively new and nascent form of computing, using high-dimensional vectors (lists of numbers that present information that can be combined in different ways to analyze the relationships between different vectors) to represent information rather than the traditional binary system.

HDC is also able to capture more complex data patterns. It is an emerging field with potential applications in machine learning (ML), natural language processing (NLP) and robotics – inspired by the patterns of neural activity in the human brain. This could allow AI-based computing systems to retain memory, which would therefore reduce their computing and power demands.

How it works: Two key concepts

Hyperdimensional computing leverages binding and superposition to simplify the analysis. Binding is a process that combines different features together to create something that encodes all of them at the same time. Superposition is a process which combines hyperdimensional vectors (hypervectors) to create a new representation that shows the relationship between the original vectors. In short, vectors can represent information just by small changes to them, and vectors can also combine to

represent new concepts and then be pulled apart again to discern how they are formed. These vectors can code info without having to add more nodes to the network.

Advantages of HDC

- **Processes vast amounts of information quickly:** HDC allows vast amounts of information to be processed concisely, which reduces memory requirements and enables more efficient storage and information retrieval. This efficiency is useful where computational resources are limited or when applications involve large-scale data processing.

- **Tolerates errors better than ANN:** If a hypervector has errors, it is still close to the original vector. This means that any reasoning using these vectors is not meaningfully impacted. These systems can be at least 10x more tolerant of hardware faults than traditional ANNs (source: Zhang et al).

- **Transparency:** How hyperdimensional computing works means we can see why the system chose the answer that it did. This is not true of traditional neural networks.

Applications in generative AI, IoT, robots, cybersecurity, healthcare

In the context of generative AI, HDC is beneficial for fine-tuning and training LLMs. It can process complex data in an IoT context and provide a unified representation of the sensor data. Further, this, combined with robots, can help them perceive and interpret sensory data about their surroundings and make decisions or even process medical sensor data to detect specific health conditions. Applications also extend to cybersecurity where it could identify patterns of cyberattacks that would typically go unnoticed with traditional approaches.

5) Artificial General Intelligence:

Artificial general intelligence (AGI) is a field of AI that attempts to create software with human-like, or 'above intelligence,' which can self-learn. It is a hypothetical type of 'AI agent' (programs that could perform actions in an iterative process to set a policy or goal, using external tools or AI models to achieve them).

Put another way, current AI technologies perform within a set of pre-determined parameters but cannot do other tasks without being programmed to do so, e.g., image-generating models can't write an essay. AGI is the field to develop systems that have self-control, self-understanding and learn new skills. With these skill sets, they can solve complex problems that they may not have been able to do when they were first created. Human-level AGI is a theoretical concept.

How are AI and AGI different? AI allows software to solve novel and difficult tasks at a level that a human can do. AGI can solve problems in different domains without manual intervention because it can self-learn. In this way, AGI is a theoretical concept of AI, which solves complex tasks with generalized human-level abilities. This also links in with the terms 'strong' and 'weak' AI. Strong AI is AGI that can perform tasks with human cognitive levels with little background knowledge, whereas weak AI includes systems limited to specific tasks they are designed for. Generative AI refers to deep-learning models that can generate high-quality content, based on the data they were trained on. The ability of an AI system to generate content does not mean its intelligence is general.

LLMs are 'intelligent' but it depends on how we define intelligence

Based on various metrics, LLMs can 'outsmart' humans on various areas of knowledge, but it depends on how we define and measure intelligence. Intelligence is a complex construct that any test/metric cannot fully capture. There have been traditional IQ and Turing tests that have been used to evaluate AI model performance. Human IQ tests probe cognitive abilities, e.g., memory, attention and problem-solving, whereas AI model evaluation assesses an AI system's performance on specific tasks or problems.

It is hard to define intelligence scientifically. However, there are three important canonical abilities to look out for as a framework: 1) reasoning; 2) planning; and 3) learning from experience. And, most crucially, not only are these three capabilities needed, but they shouldn't be limited to any specific domain.

Not at AGI yet; hallucination & planning capabilities need addressing

At a recent BofA Global Research Transforming World event, an expert speaker noted that large language models are still far from human-level intelligence and there are improvements to be made before reaching this. Hallucination and planning capabilities (the process of using autonomous techniques to solve planning and scheduling problems) could be two things to overcome first to reach AGI.

Take for instance, building a house; there is an order to the necessary steps, e.g., the foundations first then the walls. GPT models (neural network-based language prediction models built on the Transformer architecture) are currently very linear and cannot dismantle the problem to determine the order of each stage in a process. Planning could naturally emerge with newer versions like GPT-5, but alternatively, planning capabilities could be something that needs to be integrated into these models directly.

What would it take to turn AI into AGI?

There are several capabilities that are characteristic of AGI systems:[3]

- **Sensory perception:** AI systems still do not have human-like sensory perception capabilities, e.g., colour detection, determining spatial characteristics from sound, etc.
- **Fine motor skills:** dexterity to do everyday things that the average human can do. e.g., finding a set of keys.
- **Natural language understanding & problem solving:** full comprehension of books, articles, videos and common-sense knowledge to operate in the real world, e.g., recognising that a light bulb is blown and needs changing.
- **Navigation:** leveraging GPS (global positioning system) or projecting actions through imagined physical spaces.
- **Social and emotional engagement:** humans must want to interact with the AI system, not fear them. To be able to interact with humans, robots need to understand humans, interpret facial expressions or changes in tone.

It's very costly to develop LLMs because of the GPU (graphics processing unit) hardware requirements. The transition to SLMs (small language models) could be the next frontier for AI. The field is moving from "Attention Is All You Need" to "Scale Is All You Need" to "Textbooks Are All You Need," where attention and size and/or scale are not the 'be all and end all' for achieving the best AI models. Teaching AI is not going to be like how we teach human beings – because a transformer is not a human brain and there is almost no connection between the two – but it should involve a textbook-like approach.

The Australian National University identified eight attributes a system must have for it to be considered AGI: logic, autonomy, resilience, integrity, morality, emotion, embodiment, and embeddedness. Embodiment and embeddedness refer to having a physical form that facilitates an understanding of the world and human behaviour including human needs and values.

Is AGI coming sooner than we think?

Since the first discussions about general AI and technological singularity by mathematician Von Neumann in the mid-20th century, scientists and technologists have repeatedly predicted the coming of human-level intelligent machines in the near term. However, with repeated failure to deliver, the industry has seen waves of investment interest and decline. Yet, this decade has seen a resurgence of interest as the growth of data computing power and technological innovation (such as improved architecture of processors) has continued to explode. Many AI experts believe that human-level AI will be developed in the next decade, and some think much sooner. Only time will tell.

[3] "What is artificial general intelligence (AGI)," March 21, 2024, McKinsey

Contributors

Vanessa Cook

Content Strategist, Bank of America Institute

Sources

Haim Israel

Equity Strategist, BofA Global Research

Felix Tran

Equity Strategist, BofA Global Research

Martyn Briggs

Equity Strategist, BofA Global Research

Lauren-Nicole Kung

Equity Strategist, BofA Global Research

Disclosures

These materials have been prepared by Bank of America Institute and are provided to you for general information purposes only. To the extent these materials reference Bank of America data, such materials are not intended to be reflective or indicative of, and should not be relied upon as, the results of operations, financial conditions or performance of Bank of America. Bank of America Institute is a think tank dedicated to uncovering powerful insights that move business and society forward. Drawing on data and resources from across the bank and the world, the Institute delivers important, original perspectives on the economy, sustainability and global transformation. Unless otherwise specifically stated, any views or opinions expressed herein are solely those of Bank of America Institute and any individual authors listed, and are not the product of the BofA Global Research department or any other department of Bank of America Corporation or its affiliates and/or subsidiaries (collectively Bank of America). The views in these materials may differ from the views and opinions expressed by the BofA Global Research department or other departments or divisions of Bank of America. Information has been obtained from sources believed to be reliable, but Bank of America does not warrant its completeness or accuracy. Views and estimates constitute our judgment as of the date of these materials and are subject to change without notice. The views expressed herein should not be construed as individual investment advice for any particular person and are not intended as recommendations of particular securities, financial instruments, strategies or banking services for a particular person. This material does not constitute an offer or an invitation by or on behalf of Bank of America to any person to buy or sell any security or financial instrument or engage in any banking service. Nothing in these materials constitutes investment, legal, accounting or tax advice.