BANK OF AMERICA INSTITUTE

BANK OF AMERICA

Transformation

AI dictionary, part 3: Resources and risks

08 July 2025

Key takeaways

- Since AI is intertwined with other technologies, greater AI adoption can lead to more innovation and lower costs in adjacent areas too. Furthermore AI is now more than just a technology - it's also key for geopolitics (sovereign AI), economies and companies.
- But with great power, comes great responsibility, and AI doesn't come without its risks. Things like hallucination and deepfakes could erode public trust in information sources.
- The final publication in our three-part series discusses AI resources and risks - from physical components of AI (semiconductors) to the tools and processes used to develop and advance the technology (training, knowledge graphs, etc.).

This publication is the final installment of our three-part AI dictionary series. Part one covers the AI basics, from exploring the common types of AI widely used today to discussing key AI technologies, including machine learning, deep learning and natural language processing (read: AI dictionary, Part 1: The basics). The second part of our series answers the question, "What comes next?" From artificial general intelligence to artificial super intelligence, we demystify the technology and look at what's ahead (read: AI dictionary, part 2: The next generation).

Deepfakes

What are they?

Deepfakes are AI-generated content (video, sound, voice recording, etc.) designed to mimic a real-life person or scene.[1] The content may be from scratch, or pre-existing content that has been altered. They can also show original content where someone is represented doing or saying something that they have not done.[2]

For example, deepfake images can be based on existing source content where one person is swapped for another. With generative AI becoming more prevalent, so can deepfakes. In fact, by 2030, there could be 10 deepfakes for every person on the planet.[3]

Deepfakes becoming harder to detect and can be used for malicious purposes

As technology improves, deepfakes are becoming harder to detect, making it difficult to determine what is real and what is fake, which poses the risk of eroding public trust in information sources. As mentioned in a previous publication, The world in 2030: Part 1, generative adversarial networks (GANs), which detect and improve any flaws in a deepfake within multiple rounds, make it harder for deepfake detectors to decode them. GANs are also used as a popular method of creating deepfakes, relying on the study of large amounts of data to "learn" how to develop new examples that mimic the real thing, with startlingly accurate results.

Unfortunately, deepfakes are often used for malicious purposes, such as spreading misinformation and disinformation, or cybercrime – whether for financial gain, social disruption or other nefarious purposes. For example, they can be used to commit fraud and access services by pretending to be someone else, or to gain access to services they wouldn't be able to access using their true identity.

Blockchain could be a way to detect deepfakes

Content can be cryptographically signed by multiple parties at the source of origin. Cryptographic hash functions (equations used to verify data validity) can be assigned to the video at the time of recording. With blockchain's immutability feature, the hash data cannot be modified once entered. Every instance of upload, download, and edit to the video can be written into a smart

[1] Leblanc, M., & Swatton, P. (2024, June 7). *What are deepfakes and how can we detect them?* The Alan Turing Institute. https://www.turing.ac.uk/blog/what-are-deepfakes-and-how-can-we-detect-them
[2] Barney, N., Wigmore, I., & Yasar, K. (2025, May 22). *What is deepfake technology?* TechTarget. https://www.techtarget.com/whatis/definition/deepfake
[3] BofA Global Research, Sensity.

contract after validation by the original parties. This creates an audit trail for the video, assures its integrity, and improves traceability. The hash data can be compared to the source at every stage. If there is any mismatch between the two datasets, it can help conclude that the content has been altered.

Hallucination

What is it?

Hallucination is when a model generates output that seems plausible but is factually incorrect, irrelevant or nonsensical.[4]

Why? AI models are trained on data, and they learn to make predictions by finding patterns in the data. But the accuracy of these predictions often depends on the quality and completeness of the training data.[5] Hence, hallucination can occur due to various factors such as overfitting (when an algorithm fits too closely to its training data and cannot make predictions from any data other than the training data), training data bias and inaccuracy, or insufficient training data.

For example, an AI model may be trained on a dataset of medical images to be able to detect cancer cells. If the dataset does not include images of healthy tissue, the AI model may incorrectly predict that the healthy tissue is cancerous.

Additionally, hallucination can result if an AI model struggles to understand real-world knowledge, physical properties, or factual information. For example, when prompted to summarize content and it includes details that are not present in the original content, or even fabricates details.

Problematic because a user may trust the output and make decisions based on it

AI hallucination can include an incorrect prediction, false positives (identifying something as a threat when it is not), and false negatives (failing to identify something as a threat when it is). This is problematic because the user may use its output or even make decisions based on its inaccurate answers.

In addition, hallucination can be a source of misinformation and disinformation.[6] Given it may be difficult to discern between what is real and fake, it poses the risk of eroding public trust in information sources. Misleading information can also cause reputational damage to stakeholders.

High-quality training data, credible sources, and specific prompting can reduce it

High-quality training data is important to ensure that the model has been trained on data that represents the real world, encompassing various scenarios and examples to cover potential edge cases. This includes ensuring the data is free from bias and errors, or even using sources that are credible.

For example, knowledge graphs (KGs), which organize data from multiple sources, capture information about the topics of interest and forge connections between them, can mitigate hallucination by acting as a structured and reliable source of interconnected information. Specific prompting can also help, where the user gives clear, detailed instructions without leaving room for interpretation.[7]

Inference

What is it?

Inference is where a trained machine learning (ML) model draws conclusions from new data.[8] An example of AI inference would be an autonomous vehicle that can recognize a stop sign on a road that it has not driven on before. Identifying the stop sign in this new context is inference.

When it comes to designing and implementing AI models, there are two key phases: in the training phase, the model looks at an existing data set to discover patterns and relationships within it. Next, in the inference phase, the trained model applies these learned patterns to create predictions, generate content or make decisions when it encounters new, previously unseen data.

Latency matters in inference

According to BofA Global Research, typically, the inference batch size is smaller and uses less precise data (8 bits of data at a time) than training (16/32 bits of data at a time), as users don't want to wait several seconds while the system is accumulating images for a large batch. Hence, *latency*, or time taken to deliver an output, is a more important factor for inference than in training.

---

[4] Google Cloud. (n.d.). *What are AI hallucinations?* Google Cloud. https://cloud.google.com/discover/what-are-ai-hallucinations
[5] Ibid.
[6] NTT DATA Group. (n.d). *Not All Hallucinations Are Bad: The Constraints and Benefits of Generative AI.* NTT DATA. https://www.nttdata.com/global/en/insights/focus/2024/not-all-hallucinations-are-bad-the-constraints-and-benefits-of-generative-ai
[7] Digital Ocean.
[8] Cloudflare. (n.d.). *AI inference vs. training: What is AI inference?* Cloudflare. https://www.cloudflare.com/learning/ai/inference-vs-training/

Inference costs have been falling

Inference costs are falling. The cost for an AI model scoring similarly in performance to OpenAI's GPT -3.5 (a 64.8 on the Massive Multitask Language Understanding benchmark), dropped from $20.00 per million tokens (unit of text) in November 2022 to just $0.07 per million tokens by October 2024.[9] This is approximately a 99.65% reduction in just under two years. And depending on the task, large language model (LLM) inference costs continue to fall anywhere from nine to 900 times per year, allowing AI to become more affordable and accessible, which in turn, drives adoption.[10]

Increasing compute at the model use (inference) stage

Increasing AI model capabilities has implications for the computing requirements at the training and inference stages.[11] Training requires as much intelligence fed into the model at the time it's created and trained, where scale matters. Inference is where you use the model to get an answer from it and make it work, requiring not just exchanging tokens but also local computation and reasoning. This shifts the compute from only at the training phase in AI model development to also the usage stage.

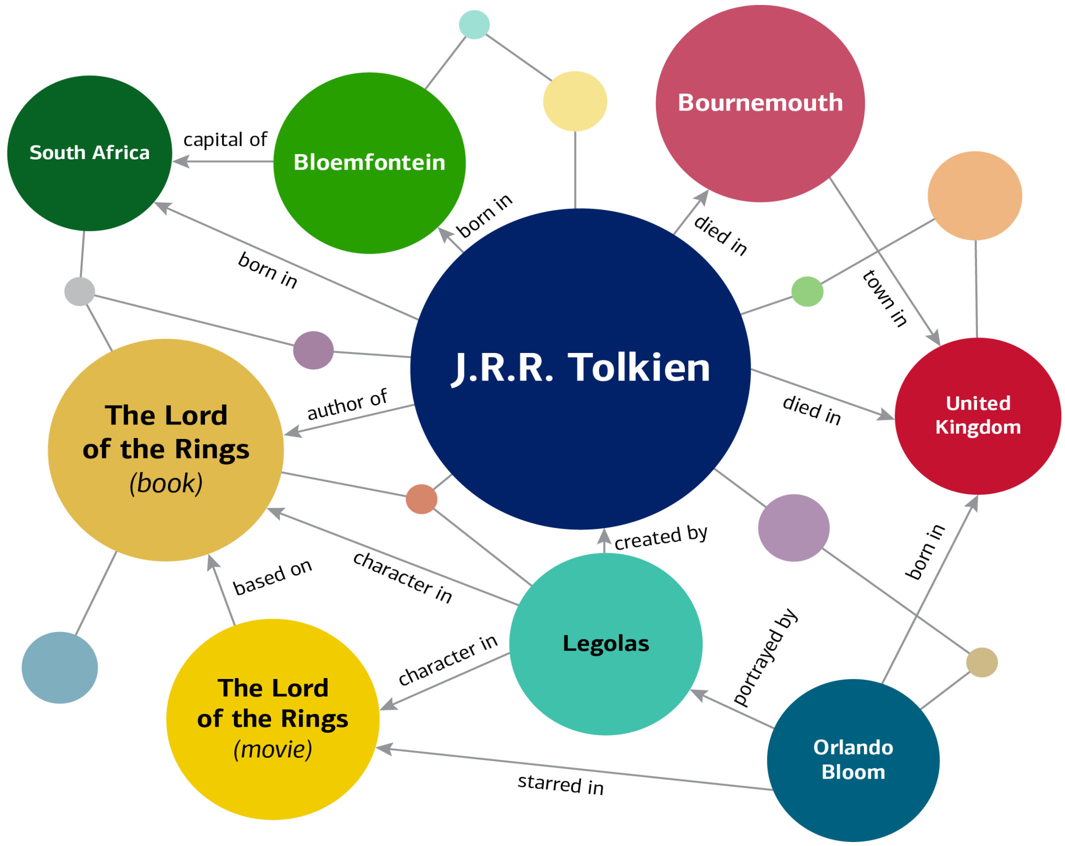
Knowledge graphs

What are they?

Knowledge graphs (KGs) organize data from multiple sources, capture information about the topics of interest, and forge connections between them, as demonstrated in Exhibit 1. KGs use machine learning (ML) and natural language processing (NLP) to construct a comprehensive view of nodes, edges and labels through a process called semantic enrichment.[12]

Exhibit 1: Knowledge graphs organize data from multiple sources, capture information about the topics of interest and forge connections between them

Sample knowledge graph on Lord of the Rings-related information



Source: Semrush, BofA Global Research

BANK OF AMERICA INSTITUTE

[9] Maslej, N. (2025, April 7). *AI Index 2025: State of AI in 10 Charts.* Stanford University. https://hai.stanford.edu/news/ai-index-2025-state-of-ai-in-10-charts

[10] Adamczewski, T., Cottier, B., Owen, D., & Snodin, B. (2025, March 12). *LLM inference prices have fallen rapidly but unequally across tasks.* Epoch AI. https://epoch.ai/data-insights/llm-inference-price-trends

[11] Steve Brown.

[12] IBM. (n.d.). *What is a knowledge graph?* IBM. https://www.ibm.com/think/topics/knowledge-graph

When data goes through the system, KGs can identify individual objects and understand the relationships between them. This knowledge is integrated with other datasets that are relevant and similar in nature. However, 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:[13]

- **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.

What are they used for?

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

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. KGs ingest huge amounts of factual information from multiple sources, forging connections between them.

Integrating a KG 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. KGs are not probabilistic engines like LLMs. Instead, they can enhance LLMs by being a centralized source of accurate knowledge for inference and interpretability, and they reduce the need for large, labelled datasets.

As an example, in the biopharmaceutical 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 KG 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 KG to identify and analyze all the information related to that compound.

One advantage of this approach is that it keeps everything accurate in one centralized place, while also making it easier to bring together information from various sources and formats.

Responsible AI

What is it?

Responsible AI is a set of principles that help guide the development and deployment of AI systems.[14] They consider AI's broad impact on society and aim for it to be deployed in a safe and ethical way.[15] Responsible AI aims to mitigate the risks and negative outcomes associated with AI, while maximizing the positive outcomes. A responsible AI framework can encompass:[16]

- **Fairness:** Developing AI systems that are equitable and ensuring that everyone is treated fairly.

- **Privacy and security:** Creating AI systems that respect user privacy and protect data from unauthorized access or misuse.

- **Explainability:** Ensuring that AI systems' decision-making processes are understandable to humans.

- **Transparency:** Understanding how AI systems have been created or how they have reached their conclusions.

[13] Horrocks, I., Jimenez, Ruiz, E., Pan, J., & Tamma, V. (n.d.). *Knowledge graphs.* The Alan Turing Institute. https://www.turing.ac.uk/research/interest-groups/knowledge-graphs

[14] Stryker, C. (2024, February 6). *What is responsible AI*? IBM. https://www.ibm.com/think/topics/responsible-ai
[15] Microsoft. (2024, September 13). *What is Responsible AI - Azure Machine Learning.* Microsoft. https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai?view=azureml-api-2
[16] Ibid., AWS, IBM, Infused Innovations

- **Governance:** Ensuring that AI systems are developed and used in a way that aligns with ethical principles, organizational values, and societal expectations.

- **Reliability and safety:** AI systems should be able to operate as they were originally designed, respond to unanticipated conditions, and resist harmful manipulation.

- **Inclusivity:** AI systems that are designed and used in a way that is accessible, usable, and beneficial to a diverse range of people.

- **Accountability:** Ensuring that individuals and organizations responsible for designing and deploying AI systems are answerable for how they operate.

Regulation also acts as a framework for developers to achieve these principles

Between 2016 and 2024, 39 out of 114 countries (over one-third) enacted at least one AI law. In total, 204 laws were passed. Exhibit 2 showcases the 15 countries who passed the greatest amount of legislation in this timeframe, with the US in the lead (27 laws). Some examples of legislation passed into law in 2024 include:[17]

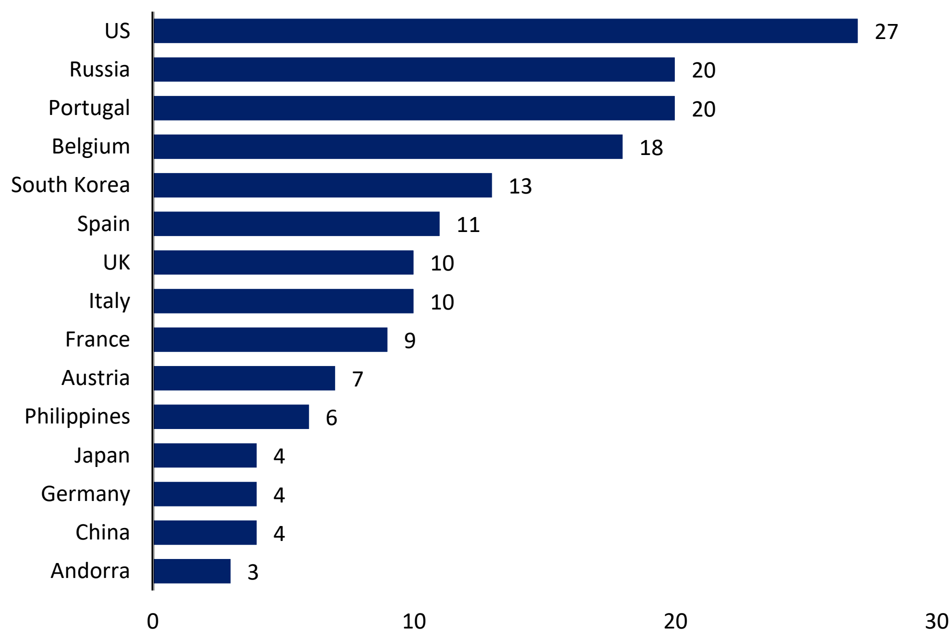
- **Austria's Federal law amending the KommAustria Act and the Telecommunications Act 2021**: The law establishes a Service Center for AI, to support, advise and coordinate AI governance in the media, telecommunication, and postal sectors. It also mandates an AI advisory board to monitor AI developments, advise the government, and shape national AI policy. To fund the activities under the act, €700,000 are allocated annually, with future adjustments based on inflation.

- **Belgium's Royal decree establishing an orientation committee on AI**: The decree created a federal AI steering committee to advise the government on AI-related policies. It also serves as the primary point of contact for AI governance.

- **Latvia's amendments to the pre-election Campaigning Law:** The amendments regulate the use of AI in political advertising, requiring clear disclosure for AI-generated content in paid campaign materials. It bans the use of automated systems with fake or anonymous social media profiles for election campaigns.

Exhibit 2: Between 2016-24, the US passed the most AI-related laws

Number of AI-related bills passed into law in select countries, 2016-24



Source: AI Index 2025 Annual Report by Stanford University. CC-BY-ND 4.0. Reformatted; BofA Global Research

BANK OF AMERICA INSTITUTE

[17] Stanford University. (2025). *2025 AI Report, Chapter 6: Policy and Governance.* Stanford University. https://hai.stanford.edu/assets/files/hai_ai-index-report-2025_chapter6_final.pdf

While the US did pass the greatest number of laws on AI from 2016 to 2024, these were not comprehensive federal laws or regulations. Rather, they were state-level legislation. Examples include:

- **California's AI Transparency Act:** This act requires large AI providers to offer free AI detection tools and ensure AI-generated content includes clear, permanent disclosures. Violations result in a $5,000 fine per instance, with enforcement by the attorney general or local authorities.

- **Colorado's Consumer Protections for Artificial Intelligence:** This law establishes consumer protections for interactions with high-risk AI systems. It requires the developers and deployers to prevent algorithmic discrimination. The AI systems must provide transparency and allow consumers to correct or appeal AI-driven decisions.

European Union (EU) AI Act

The EU AI Act is the primary legislative framework for regulating AI in the EU. It was published in the Official Journal of the European Union on July 12, 2024, and is the first comprehensive horizontal legal framework for regulating AI across the region. It entered into force on August 1, 2024, and will be effective from August 2, 2026 (except for specific provisions listed in Article 113).

The EU AI Act applies to all sectors, and categorizes AI systems according to risk (Exhibit 3):[18]

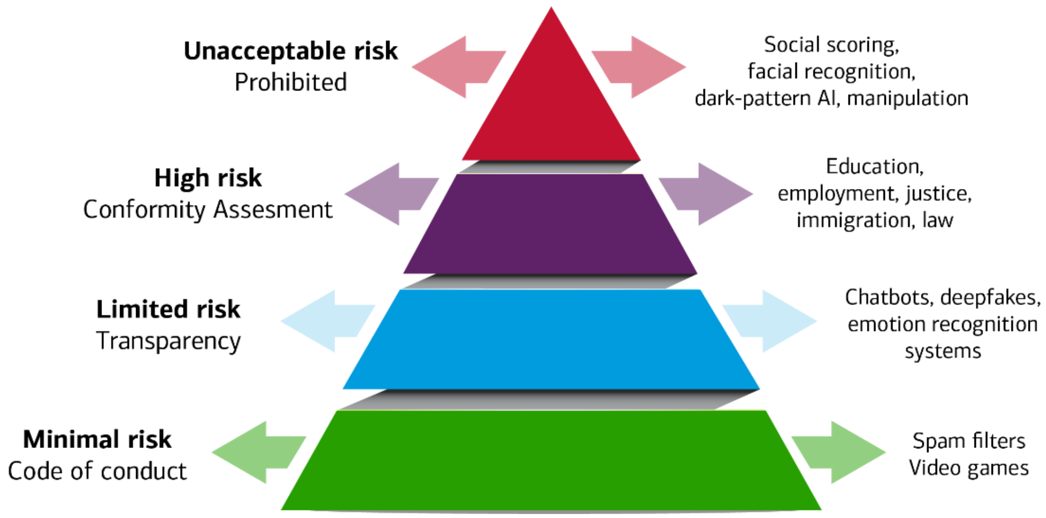
- **Unacceptable risk:** Models with unacceptable risk are prohibited. This includes AI systems used for social scoring and which are deceptive or use exploitative techniques to materially distort a person's behavior in a manner that can cause harm.

- **High risk:** Includes AI systems that fall within two categories: 1) used as a safety component or a product; or 2) deployed in eight specific areas, including education, employment, access to essential public and private services, law enforcement, migration, and the administration of justice.

- **Limited risk:** AI systems that directly interact with natural persons (e.g., chatbots), emotion recognition systems, biometric categorization systems, and AI systems that generate deepfakes. These systems are required to disclose whether the content has been artificially generated or manipulated.

- **Low or minimal risk:** Any AI system that is not caught by the above.

Exhibit 3: The EU AI Act categorizes AI systems into four levels of risk: unacceptable, high, limited and minimal
Illustrating the four risk levels identified under the EU AI Act



Source: Ada Lovelace Institute, Lilian Edwards. CC-BY-4.0. Amended text; BofA Global Research

BANK OF AMERICA INSTITUTE

[18] White & Case. (2025, April 11). AI Watch: Global regulatory tracker - European Union | White & Case LLP. White & Case. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-european-union

Retrieval augmented generation

What is it?

Retrieval augmented generation (RAG) is a technique for enhancing the accuracy and reliability of generative AI models with information from specific and relevant data sources.[19] It involves connecting an external data source to a generative AI model to create responses.

RAGs operate via two steps. In the retrieval and pre-processing stage, RAG leverages search algorithms to query external data, such as webpages and databases. Once the information is retrieved, the relevant information undergoes pre-processing, including tokenization (breaking down text into digestible units). In the grounded generation stage, the pre-processed retrieved information is incorporated into the pre-trained model.[20] The integration enables the model's output to be more precise.

Access to latest information, factual grounding for generative AI, faster data retrieval

RAG has several advantages for augmenting traditional methods of text generation, especially when dealing with factual information or data-driven responses. For example, it can provide access to fresh information, which is beneficial when some generative AI models are limited to their pre-trained data. RAG can also be a source of factual grounding, which is beneficial if a model's training data contains inaccuracies or biases.[21]

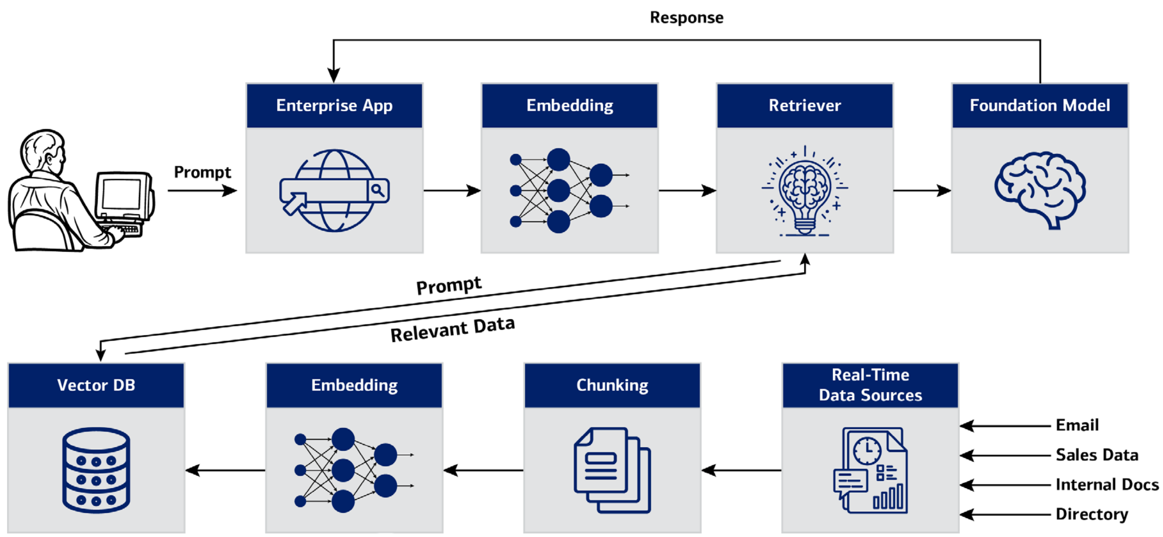
Furthermore, RAG can be integrated into any generative AI application or agent that needs access to data, such as chatbots and conversational agents.[22] In fact, the average information worker spends 32 days a year, or 13% of their time, looking for the information required for their job. But by leveraging RAG, users could more easily access data, and thus be more productive. Beyond workers, it can also improve access to information for customers.

Additionally, in retrieving relevant information from a knowledge base and incorporating it into the generation process, RAG can reduce hallucination and improve the accuracy of the generated output.[23]

RAG solutions may increase output accuracy, decrease usage costs and mitigate risks

Enterprise generative AI apps will likely leverage small language models (SLMs) and RAG solutions to enable deployment of enterprise-specific use cases that integrate internal, proprietary, and real-time data. For example, some companies are integrating AI-powered products into their respective cloud platforms, including customer relationship management (CRM) and contact center as a service (CCaaS) offerings that merge and leverage real-time data across numerous communication channels to identify inefficiencies and drive process optimizations. SLMs and RAG also decrease usage costs, improve contextual awareness, increase output accuracy, and reduce output latency (Exhibit 4).

Exhibit 4: Generative AI apps + RAG solutions may increase output accuracy, decrease usage costs and mitigate risks
Retrieval augmented generation (RAG) solutions enable enterprises to deploy specialized use cases



Source: BofA Global Research

BANK OF AMERICA INSTITUTE

---

[19] Merritt, R. (2025, January 31). *What Is Retrieval-Augmented Generation, aka RAG?* NVIDIA. https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/
[20] Google Cloud. (n.d.). *What Is Retrieval Augmented Generation (RAG)?* Google Cloud. https://cloud.google.com/use-cases/retrieval-augmented-generation
[21] Ibid.
[22] Ibid.
[23] Ibid.

Semiconductors

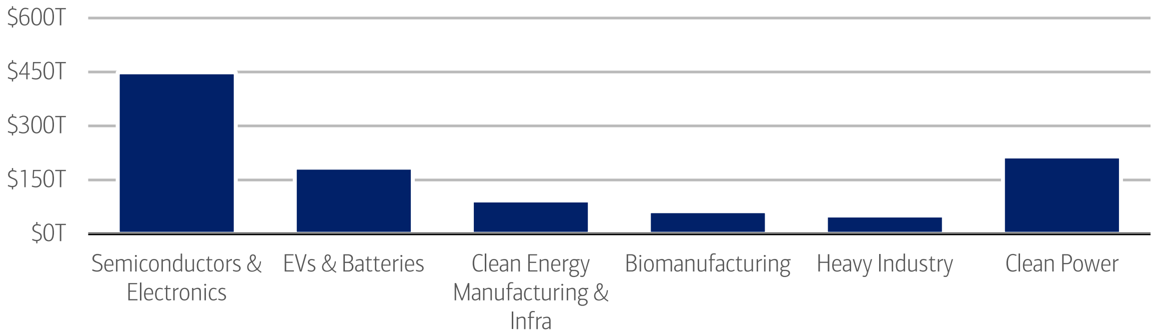
What are they?

Semiconductors (or "semis"), commonly known as "chips" (short for microchips), are integral components for virtually every electronic device. Different chips range in "intelligence" and function, but the purpose for most is to process, store, and/or transmit data. In other words, they are the "brain" of a device.

Deepening technology adoption means we need significantly more resources to enable the productivity gains and economic growth potential from AI and future technologies. One of the key resources for this is semis. AI is set to accelerate demand for more powerful chips, such as graphic processing units (GPUs), with the total available market (TAM) for AI accelerators set to triple between 2024 and 2030 to approximately $360 billion, according to BofA Global Research. Manufacturing capacity and supply chains are expanding to accommodate, with an overall semis market opportunity of over $1 trillion (Exhibit 5).[24]

Exhibit 5: Private US manufacturing investments announced since 2021 exceed $1 trillion, half of which is in semiconductors and electronics

Infrastructure investment is reviving US manufacturing ($, trillions)



Source: White House Invest in America as of January 10, 2025; BofA Global Research

BANK OF AMERICA INSTITUTE

According to BofA Global Research, examples of semis include:

- **Central processing units (CPUs):** They perform one of the most important roles in any computing device – processing data and controlling the other chips, like the "brain" of an electronic device.
- **Graphics processing units (GPUs):** These are specialized processors that can be used for AI model training and inference, due to their parallel computing capabilities. They can process billions of calculations based on datasets to identify the patterns used in AI inference.

  GPUs were originally designed to render complex images for PC video games and films. Unlike a CPU, GPUs consist of hundreds to thousands of processing cores, and each is dedicated to performing a specific operation in parallel. In recent years, engineers discovered that parallel computation could be used to handle big data applications more efficiently than just a CPU. So, the GPU has been extended into adjacent markets, such as AI, self-driving cars, workstations, cloud data centers, and high-performance computing. A GPU, or multiple GPUs, typically interface with one or more CPUs. The CPUs are responsible for general tasks like programming and data movement, while the GPU is used to handle the bulk of the complex parallel computations.
- **Application specific integrated circuits (ASICs):** Integrated circuits are designed and manufactured for the exclusive use of a single customer. The ASIC may be a digital, analog, or mixed-signal chip, and typically combines functional blocks such as processor cores, memory, interface circuits, analog, and custom logic. It's important to note that integrated circuits are singular semiconductor chips designed with many other discrete semiconductor components (resistors, capacitors, inductors, diodes) that are integrated into its geometry.
- **Field programmable gate arrays (FPGAs):** Integrated circuits that are designed to be configured for different algorithms after manufacturing – hence the term "field programmable." FPGAs are based on a matrix of configurable logic blocks (CLBs) connected to input/output blocks (IOBs) and each other via programmable interconnects. Modern FPGA devices consist of millions of logic cells that can be configured to implement a variety of software algorithms.

[24] Jong, M.d., Marcil, H., Wagner, R., & Wiseman, B. (2025, April 21). *Semiconductors have a big opportunity—but barriers to scale remain.* McKinsey & Company. https://www.mckinsey.com/industries/semiconductors/our-insights/semiconductors-have-a-big-opportunity-but-barriers-to-scale-remain

Sovereign AI

What is it?

Sovereign AI refers to a nation's capabilities to produce AI using its own infrastructure and resources.[25] And there's a pressing reason why countries should try to keep up – with c. $16 trillion annual global economic value on the line, AI is becoming key in setting the new world order and nations at the forefront of the revolution can gain significant strategic advantage. It is not just the race for technological development but also for resources, supply chains, regulation and standards that are on the line. Whoever controls AI could have an advantage, possibly reshaping the geopolitical balance of power.

As such, nations around the world are investing in sovereign AI. For example, in February 2024, Singapore's Deputy Prime Minister and Finance Minister announced the government's plans to allocate more than US$1 billion over the next five years to support AI, talent development and industry growth.[26] In March 2024, India launched the IndiaAI Mission to strengthen its AI ecosystem. The US$1.25 billion initiative aims to build over 10,000 GPUs through public-private partnerships and support domestic AI models.[27]

What can impact a country's ability to build sovereign AI?

As discussed in [World in 2030: Part 1](), technology is moving us to a new phase in the next five years. Powered by the AI revolution, BofA Global Research believes that we will watch technology prices plummet and see AI's integration in all aspects of our lives while witnessing its game-changing role in leap-frogging innovation. As discussed in part two of this series, Agentic AI will influence the job market, and rich AI simulations will develop new products in healthcare, industrials and financial services.

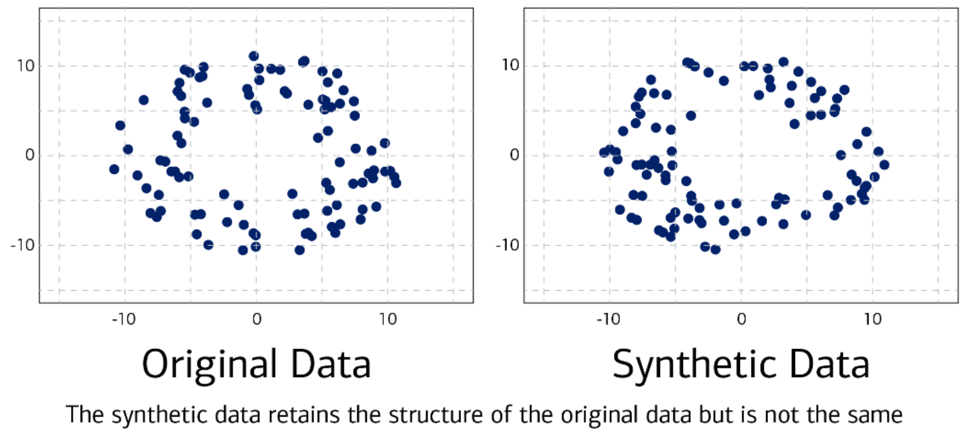
Furthermore, AI will interact with the physical environment, enabling the next generation of automation. At the same time, we are likely to see a tech war "arms race" between the superpowers, complicated by accelerated deglobalization and tech protectionism, as well as privacy and demographic concerns. Therefore, access to different resources needed for AI such as computing power, energy, data, metals, water, bandwidth, and talent can affect the path to achieving sovereign AI.

Synthetic Data

What is it?

Synthetic data is artificial data designed to mimic real world data. It is created using statistical methods (e.g. deep learning (DL) and generative AI). When created by generative AI models, the algorithms learn the statistical properties of the sample data and then the model can create statistically identical synthetic data (Exhibit 6), which allows it to supplement or replace real data sets.

Exhibit 6: Synthetic data is artificial data that retains the underlying statistical properties of the original data on which it is based, allowing it to supplement or even replace real data sets
Illustrating the difference between original and synthetic data



Source: Data in Government. Public sector information licensed under the Open Government Licence v3.0; BofA Global Research
BANK OF AMERICA INSTITUTE

[25] Lee, A. (2024, February 28). *What Is Sovereign AI?* NVIDIA. https://blogs.nvidia.com/blog/what-is-sovereign-ai/
[26] Stanford University. (2025). *2025 AI Report, Chapter 6: Policy and Governance.* Stanford University. https://hai.stanford.edu/assets/files/hai_ai-index-report-2025_chapter6_final.pdf
[27] Ibid.

Synthetic data can be used to train models where data availability is limited

Synthetic data can be used to train machine learning (ML) models, particularly in areas where there is limited data availability, or where it can be difficult or time-consuming to obtain real world data.[28]

In the case of training embodied or physical AI, like autonomous vehicles or humanoid robots, it may be difficult and time-consuming to obtain data on all the real-world scenarios. For example, if there is a particular driving turn too complex to replicate in the real world, synthetic data simulation could be used instead, and input directly into the car's mainframe computer. Even the pharmaceutical industry could leverage synthetic data. Instead of physically carrying out many drug trials over several years, companies could create a simulated population of the demographic groups of interest to test their drugs in a shorter timeframe.

Training

What is it?

Training involves teaching an AI model to recognize patterns and make predictions based on data. To fully train the model, the system needs to backpropagate (feedback) errors and adjust the weights of neurons (interconnected nodes) accordingly – this is similar to how children learn to identify objects. Typically, to train a neural network, one million to one billion parameters must be adjusted, and simultaneous operation is important to minimize time of operation.

Throughput matters in the training process

BofA Global Research notes that in training, a data scientist takes data and sends it through a model; this model continuously iterates and gets feedback on its accuracy, which requires a large amounts of computing power. Throughput, or the number of operations executed per second, is a key attribute for training a neural network. Much of the initial work in developing and training neural networks has been carried out using graphic processing units (GPUs) as they are the processor capable of executing billions of operations per second in parallel.

As AI models grow, more training compute and energy are required

Larger model size translates into greater computational intensity to train the models. Epoch estimates that the training compute of notable AI models doubles roughly every five months and large language model (LLM) training datasets double every eight months. This rapid rise in compute demand has important implications. For instance, models requiring more computation often have larger environmental footprints. In fact, the power required for training doubles every year.[29]

---

[28] IBM. (2023, January 31). *What is synthetic data?* IBM. https://www.ibm.com/think/topics/synthetic-data, Mostly AI

[29] Emberson, L., & Rahman, R. (2024, September 19). *The power required to train frontier AI models is doubling annually*. Epoch AI. https://epoch.ai/data-insights/power-usage-trend

Contributors

Vanessa Cook

Content Strategist, Bank of America Institute

Lynelle Huskey

Analyst, Bank of America Institute

Sources

Haim Israel

Equity Strategist, BofA Global Research

Lauren-Nicole Kung

Equity Strategist, BofA Global Research

Felix Tran

Equity Strategist, BofA Global Research

Martyn Briggs

Equity Strategist, BofA Global Research

Disclosures