Hype Cycle for Artificial Intelligence, 2023

Published 19 July 2023 - ID G00791179 - 123 min read

By Analyst(s): Afraz Jaffri

Initiatives: Artificial Intelligence; Evolve Technology and Process Capabilities to Support D&A; Generative AI Resource Center

Generative artificial intelligence and ChatGPT have elevated AI discussions to new levels, pushing interest to boardrooms and heads of state alike. Data and analytics leaders must stay on top of the trends and track the trajectory of innovations to create credible cases for investment.

Additional Perspectives

- Summary Translation: Hype Cycle for Artificial Intelligence, 2023
  (24 August 2023)

More on This Topic
This is part of an in-depth collection of research. See the collection:

- 2023 Hype Cycles: Deglobalization, AI at the Cusp and Operational Sustainability

Analysis

What You Need to Know

Generative AI has had an impact like no other technology in the past decade. The increased productivity for developers and knowledge workers, using systems like ChatGPT, is very real and has caused organizations and industries to rethink their business processes and the value of human resources.

In turn, the apparent abilities of generative AI systems have rekindled debates on the safe usage of AI and whether artificial general intelligence can be achieved, or has even already arrived. Current generative AI techniques are fallible, however, and many of the innovations on this year's Hype Cycle need to be put together in order to go beyond the limitations and mitigate the risks.

Data and analytics (D&A) leaders must leverage this research to prepare their AI strategy for the future and utilize technologies that offer high impact in the present.

The Hype Cycle

Generative AI is dominating discussions on AI and has reached the Peak of Inflated Expectations, together with foundation models, which have become bigger and exhibit behaviors that display human-level performance on a variety of complex comprehension tasks. The hype surrounding these models, often in news and media, generally focuses on cherry-picked examples of output rather than a realistic assessment of their strengths and weaknesses. There is still a large gap between the expected potential impact and actual usage,

AI-related innovations that have moved past the peak and are entering the Trough of Disillusionment include synthetic data, edge AI, ModelOps and knowledge graphs. Knowledge graphs are the second biggest mover on the Hype Cycle and have been touted as the solution to many of the problems with generative AI techniques, but still require some work to become a mainstream technology.

Intelligent applications, cloud AI services, data labeling and annotation, and computer vision are moving toward the Plateau of Productivity, all also expedited by generative AI advancements.

To summarize, Gartner sees two sides to the generative AI movement on the path toward more powerful AI systems:

The innovations that will be fueled by generative AI:

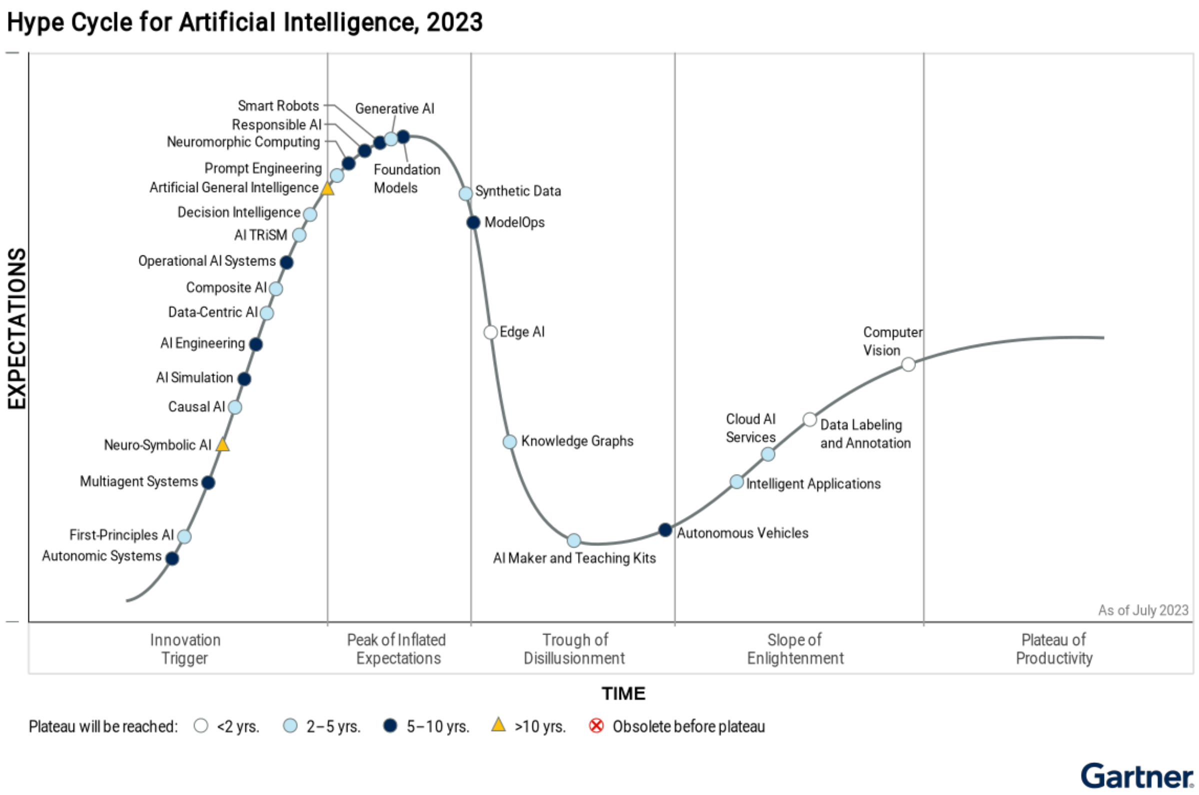
- Autonomic systems
- AI engineering
- Data-centric AI
- Composite AI
- Operational AI systems
- AGI
- Prompt Engineering
- Smart Robots
- ModelOps

- Edge AI
- Synthetic data
- Intelligent applications
- Cloud AI services
- Computer vision

The innovations that will fuel generative AI advancement:

- First-principles AI
- Neuro-symbolic AI
- Multiagent systems
- Causal AI
- AI simulation
- AI TRiSM
- Responsible AI
- Foundation models
- Knowledge graphs
- Data labeling and annotation

Figure 1: Hype Cycle for Artificial Intelligence, 2023



The Priority Matrix

Compared with other Hype Cycles, the AI Hype Cycle has more innovations with benefit ratings in the high to transformational categories, with no innovation having a benefit rating of low or moderate.

Those innovations that deserve particular attention within the two- to five-year period to mainstream adoption include generative AI and decision intelligence. Early adoption of these innovations will lead to significant competitive advantage and ease the problems associated with utilizing AI models within business processes.

Several innovations have a five- to 10-year period to mainstream adoption, and from these, responsible AI and foundational models should already be applied with small-scale projects to deliver immediate impact.

D&A leaders should balance the strategic exploration of high-value propositions with those that do not require extensive engineering or data science proficiency, and that have been commoditized as stand-alone applications and within packaged business solutions. These innovations include computer vision, knowledge graphs, smart robots, intelligent applications and AI cloud services.

Table 1: Priority Matrix for Artificial Intelligence, 2023

(Enlarged table in Appendix)

Benefit	*Years to Mainstream Adoption*			
↓	***Less Than 2 Years*** ↓	***2 - 5 Years*** ↓	***5 - 10 Years*** ↓	***More Than 10 Years*** ↓
Transformational	Computer Vision	Composite AI Decision Intelligence First-Principles AI Generative AI Intelligent Applications	Autonomic Systems Autonomous Vehicles Foundation Models Neuromorphic Computing Responsible AI	Artificial General Intelligence
High	Data Labeling and Annotation Edge AI	AI Maker and Teaching Kits AI TRiSM Causal AI Cloud AI Services Data-Centric AI Knowledge Graphs Prompt Engineering Synthetic Data	AI Engineering AI Simulation ModelOps Multiagent Systems Operational AI Systems Smart Robots	Neuro-Symbolic AI
Moderate				
Low				

Source: Gartner (July 2023)

Off the Hype Cycle

The following innovations were dropped from this year's Hype Cycle:

- **Natural language processing:** This has been broken down into several technologies, covered as part of Hype Cycle for Natural Language Technologies, 2023.
- **Digital ethics:** This has been subsumed into responsible AI, which appears on this Hype Cycle.
- **Deep learning:** The AI Hype Cycle provides a view of technologies at a high level. Specific machine learning techniques are covered as part of Hype Cycle for Data Science and Machine Learning, 2023.

On the Rise

Autonomic Systems

Analysis By: Erick Brethenoux, Nick Jones, David Cearley

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Autonomic systems are self-managing physical or software systems, performing domain-bounded tasks, that exhibit three fundamental characteristics: autonomy (execute their own decisions and tasks autonomously without external assistance); learning (modify their behavior and internal operations based on experience, changing conditions or goals); agency (have a sense of their own internal state and purpose that guides how and what they learn and enables them to act independently).

Why This Is Important

Autonomic systems are emerging as an important trend as they enable levels of business adaptability, flexibility and agility that can't be achieved with traditional AI techniques alone. Their flexibility is valuable in situations where the operating environment is unknown or unpredictable, and real-time monitoring and control aren't practical. Their learning ability is valuable in situations where a task can be learned even though there is no well-understood algorithm to implement it.

Business Impact

Autonomic systems excel where:

- Conventional automation applying composite AI techniques is inadequate, or using fixed training data is impractical or not agile.
- It is impractical to provide real-time human guidance, or training conditions can't be anticipated.
- We cannot program the exact learning algorithm, but the task is continuously learnable.

- Continuously or rapidly changing tasks or environments make frequent retraining and testing of ML systems too slow or costly.

Drivers

Autonomic systems are the culmination of a three-part trend:

- Automated systems are a very mature concept. They perform well-defined tasks and have fixed deterministic behavior (e.g., an assembly robot welding cars). The Increasing number of use cases around automation using AI techniques is a strong base for autonomous systems.
- Autonomous systems go beyond simple automation to add independent behavior. They may exhibit some degree of adaptive behavior, but are predominantly under algorithmic control (e.g., self-driving cars or a Boston Dynamics' Spot robot 1 that has its overall route and goals set by a remote human operator but has substantial local autonomy over how it achieves them). Adaptive AI capabilities are a necessary foundation for autonomic systems and should accelerate the adoption of autonomic systems.
- Autonomic systems exhibit adaptive behavior through learning and self-modifying algorithms (e.g., Ericsson has demonstrated the use of reinforcement learning and digital twins to create an autonomic system that dynamically optimizes 5G network performance. It learns from network behavior and local conditions and adjusts software and physical network control parameters to optimize performance). This trend is showing the feasibility of such systems and early learning about carefully bounded autonomic systems will build trust in their capabilities to operate independently.

Longer-term drivers include:

- Autonomic behavior is a spectrum. For example, chatbots learn from internet discussions; streaming services learn which content you like; delivery robots share information about paths and obstructions to optimize fleet routes. The advantages of systems that can learn and adapt their behavior will be compelling, and many examples will involve physical devices.
- Substantial academic research is underway on autonomics, which will result in more widespread use.

Obstacles

- **Nondeterminism:** Systems that continuously learn and adapt their behavior aren't predictable. This will pose challenges for employees and customers who may not understand how and why a system performed as it did.
- **Immaturity:** Skills in the area will be lacking until autonomics becomes more mainstream. New types of professional services may be required.
- **Social concerns:** Misbehavior, nondeterminism or lack of understanding could generate public resistance when systems interact with people.
- **Digital ethics and safety:** Autonomic systems will require architectures and guardrails to prevent them from learning undesirable, dangerous, unethical or even illegal behavior when no human is validating the system.
- **Legal liability:** It may be difficult for the supplier of an autonomic system to take total responsibility for its behavior because that will depend on the goals it has set, its operating conditions and what it learned.

User Recommendations

- Start by building experience with autonomous systems first to understand the constraints and requirements (legal, technical and cultural) that the organization is subjected to. Pilot autonomic technologies in cases where early adoption will deliver agility and performance benefits in software or physical systems.
- Manage risk in autonomic system deployments by analyzing the business, legal and ethical consequences of deploying autonomic systems — which are partially nondeterministic. Do so by creating a multidisciplinary task force.
- Optimize the benefits of autonomic technologies by piloting them in situations, such as complex and rapidly changing environments where early adoption will deliver agility and performance benefits in either software or physical systems.

Sample Vendors

Adapdix; IBM; Latent AI; Playtika; Vanti

Gartner Recommended Reading

Top Strategic Technology Trends for 2022: Autonomic Systems

First-Principles AI

Analysis By: Erick Brethenoux, Svetlana Sicular

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

First-principles AI (FPAI) (aka physics-informed AI) incorporates physical and analog principles, governing laws and domain knowledge into AI models. In contrast, purely digital AI models do not necessarily obey the fundamental governing laws of physical systems and first principles — nor generalize well to scenarios on which they have not been trained. FPAI extends AI engineering to complex system engineering and model-based systems (like agent-based systems).

Why This Is Important

As AI expands in engineering and scientific use cases, it needs a stronger ability to model problems and better represent their context. Digital-only AI solutions cannot generalize well enough beyond training, limiting their adaptability. FPAI instills a more reliable representation of the context and the physical reality, yielding more adaptive systems. A better ability to abstract leads to reduced training time, improved data efficiency, better generalization and greater physical consistency.

Business Impact

Physically consistent and scientifically sound AI models can significantly improve applicability, especially in engineering use cases (using IoT data). FPAI helps train models with fewer data points and accelerates the training process, helping models converge faster to optimal solutions. It improves the generalizability of models to make reliable predictions for unseen scenarios, including applicability to nonstationary systems, as well as enhances transparency and interpretability to make models more trustworthy.

Drivers

- **FPAI approaches instill a more flexible representation of the context and conditions in which systems operate, allowing software developers to build more adaptive systems.** Traditional business modeling approaches have been brittle. This is because the digital building blocks making up solutions cannot generalize well enough beyond their initial training data, therefore limiting the adaptability of those solutions.
- **FPAI approaches provide additional physical knowledge representations, such as partial differential equations to guide or bound AI models.** Traditional AI techniques, particularly in the machine learning family, have been confronted with severe limitations — especially for causality and dependency analysis, admissible values, context flexibility and memory retention mechanisms. Asset-centric industries have already started leveraging FPAI in physical prototyping, predictive maintenance or composite materials analysis, in conjunction with augmented reality implementations.
- **Complex systems like climate models, large-scale digital twins and complex health science problems are particularly challenging to model.** Composite AI approaches provide more concrete answers and manageable solutions to these problems, but their engineering remains a significant challenge. FPAI provides more immediate answers to these problems.
- **First principles knowledge simplify and enrich AI approaches** by defining problem and solution boundaries, reducing the scope of traditionally brute force approach employed by ML; for example, known trajectories of physical objects simplify AI-enabled sky monitoring. First-principles-based semantics reveal deepfakes.
- **The need for more robust and adaptable business simulation systems will also promote the adoption of FPAI approaches.** With a better range of context modelization and more accurate knowledge representation techniques, simulations will be more reliable and account for a wider range of possible scenarios — all better anchored in reality.

Obstacles

- From a diagnostic perspective, the development of systematic tests and standardized evaluation for these models across benchmark datasets and problems could slow down the adoption of FPAI capabilities.
- Computationally, the scaling of the training, testing and deployment of complex FPAI models on large datasets in an efficient manner will also be an issue.
- Resourcewise, collaboration across many diverse communities (physicists, mathematicians, computer scientists, statisticians, AI experts and domain scientists) will also be a challenge.
- Brute force approach is prevalent in AI, and is easy to implement for data scientists, while first principles require additional fundamental knowledge of a subject, calling for a multidisciplinary team.

User Recommendations

- **Set realistic development objectives** by identifying errors that cannot be reduced and discrepancies that cannot be addressed, including data quality.
- **Encourage reproducible and verifiable models** starting with small-scoped problems; complex systems (in the scientific sense of the term) are generally good candidates for this approach.
- **Enforce standards for testing accuracy and physical consistency** for physics and first-principles-based models of the relevant domain, while characterizing sources of uncertainty.
- **Promote model-consistent training** for FPAI models and train models with data characteristics representative of the downstream application, such as noise, sparsity and incompleteness.
- **Quantify generalizability in terms of how performance degrades** with degree of extrapolation to unseen initial and boundary conditions and scenarios.
- **Ensure relevant roles and education** in a multidisciplinary AI team (with domain expertise), so the team can develop the most effective and verifiable solution.

Sample Vendors

Abzu; IntelliSense.io; MathWorks; NNAISENSE; NVIDIA

Gartner Recommended Reading

Innovation Insight: AI Simulation

Innovation Insight for Composite AI

Go Beyond Machine Learning and Leverage Other AI Approaches

Innovation Insight: Causal AI

Predicts 2023: Simulation Combined With Advanced AI Techniques Will Drive Future AI Investments

Multiagent Systems

Analysis By: Leinar Ramos, Anthony Mullen, Pieter den Hamer

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

A multiagent system (MAS) is a type of AI system composed of multiple, independent (but interactive) agents, each capable of perceiving their environment and taking actions. Agents can be AI models, software programs, robots and other computational entities. Multiple agents can work toward a common goal that goes beyond the ability of individual agents, with increased adaptability and robustness.

Why This Is Important

Current AI is focused on the creation of individual agents built for specific use cases, limiting the potential business value of AI to simpler problems that can be solved by single monolithic models. The combined application of multiple autonomous agents can tackle complex tasks that individual agents cannot, while creating more adaptable, scalable and robust solutions. It is also able to succeed in environments where decentralized decision making is required.

Business Impact

Multiagent systems can be used in:

- **Robotics:** Swarms of robots and drones for warehouse optimization, search and rescue, environment monitoring, and other use cases
- **Generative AI:** Orchestrating large language model agents for complex tasks
- **Energy and utilities:** Smart grid optimization and load balancing
- **Supply Chain:** Optimizing scheduling, planning, routing, traffic signal control and supply chain optimization
- **Telecom:** Network optimization and fault detection
- **Healthcare:** Using agents to model actors (individuals, households, professionals) in a complex healthcare ecosystem

Drivers

- **Generative AI:** Large language models (LLMs) are increasingly augmented with additional capabilities, such as tools and internal memory, to make them better agents. Assembling and combining different LLM-based agents is increasing the interest in multiagent systems.
- **Increased decision making complexity:** AI is increasingly used in real-world engineering problems containing complex systems, where large networks of interacting parts exhibit emergent behavior that cannot be easily predicted. The decentralized nature of multiagent systems makes them more resilient and adaptable to complex decision making.
- **Simulation and multiagent reinforcement learning:** Advances in the realism and performance of simulation engines, as well as the use of new multiagent reinforcement learning techniques, allow for the training of multiagent AI systems in simulation environments, which can then be deployed in the real world.

Obstacles

- **Training complexity:** Multiagent systems are typically harder to train and build than individual agents. These systems can exhibit emergent behavior that is hard to predict in advance, which increases the need for robust training and testing. There might be, for example, conflicting objectives and interactions between agents that create undesirable behavior.
- **Limited adoption and readiness:** Despite its benefits, the application of multiagent systems to real-world problems is not yet widespread, which creates a lack of enterprise awareness and readiness to implement. Business partners might struggle to understand why a multiagent simulation is required.
- **Specialized skills required:** Building and deploying multiagent systems requires specialized skills beyond traditional AI skills. In particular, the use of reinforcement learning and simulation environments are typically required.
- **Fragmented vendor landscape:** A fragmented vendor landscape inhibits customer adoption and engagement.

User Recommendations

- Use multiagent systems for complex problems that require decentralized decision making and cannot be solved by single-agent AI models. This includes problems with changing environments where agents need to adapt and problems where a diverse set of agents with different expertise can be combined to accomplish a goal.
- Invest in the use of simulation technologies for AI training, as simulation is the primary environment to build and test multiagent systems.
- Identify areas in your organization where multiagent systems can provide an advantage over single-agent AI development. Prioritize initial experiments based on feasibility and potential value.
- Educate your data science teams on multiagent systems, how this differs from single-agent AI design, and what are some of the available techniques to train and build these systems, such as reinforcement learning.

Sample Vendors

Ansys; Cosmo Tech; FLAME GPU; MathWorks; The AnyLogic Company

Gartner Recommended Reading

Innovation Insight for Composite AI

Innovation Insight: AI Simulation

Neuro-Symbolic AI

Analysis By: Erick Brethenoux, Afraz Jaffri

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Neuro-symbolic AI is a form of composite AI that combines machine learning methods and symbolic systems (for example, knowledge graphs) to create more robust and trustworthy AI models. This combination allows statistical patterns to be combined with explicitly defined rules and knowledge to give AI systems the ability to better represent, reason and generalize concepts. This approach provides a reasoning infrastructure for solving a wider range of business problems more effectively.

Why This Is Important

Neuro-symbolic AI is important because it addresses limitations in current AI systems, such as incorrect output (hallucinations in large language models [LLMs]), generalization to a variety of tasks and explaining the steps that led to an output. This leads to more powerful, versatile and interpretable AI solutions and allows AI systems to tackle more complex tasks with humanlike reasoning.

Business Impact

Neuro-symbolic AI will have an impact on the efficiency, adaptability and reliability of AI systems used across business processes. The integration of logic and multiple reasoning mechanisms brings down the need for ever larger ML models and their supporting infrastructure. These systems will rely less on the processing of huge amounts of data, making AI agile and resilient. Decision making can be augmented and automated using neuro-symbolic approaches with less risk of unintended consequences.

Drivers

- Limitations of AI models that rely purely on machine learning techniques that focus on correlation over understanding and reasoning. The newest generation of LLMs is well-known for its tendency to give factually incorrect answers or produce unexpected results.
- The need for explanation and interpretability of AI outputs that are especially important in the regulated industry use cases and in systems that use private data.
- The need to move toward semantics over syntax in systems that deal with real-world entities in order to ground meaning to words and terms in specific domains.
- The set of tools available to combine different types of AI models is increasing and becoming easier to use for developers, data scientists and end users. The dominant approach is to chain together results from different models (composite AI) rather than using single models that are neuro-symbolic in nature.
- The integration of multiple reasoning mechanisms necessary to provide agile AI systems eventually leading to adaptive AI systems.

Obstacles

- Most neuro-symbolic AI methods and techniques are being developed in academia or industry research labs. Despite the increase in tools available, there are still limited implementations in business or enterprise settings.
- There are no agreed-upon techniques for implementing neuro-symbolic AI and disagreements continue between researchers and practitioners on the effectiveness of combining approaches; despite the emergence of real-world use cases.
- The commercial and investment trajectory for AI startups allocates almost all capital to deep learning approaches leaving only those willing to bet on the future to invest in neuro-symbolic AI development.
- Popular media and academic conferences do not give as much exposure to the neuro-symbolic AI movement as compared to other approaches, for now.

User Recommendations

- Adopt composite AI approaches when building AI systems by utilizing a range of techniques that increase the robustness and reliability of AI models. Neuro-symbolic AI approaches will fit into a composite AI architecture.
- Dedicate time to learning and applying neuro-symbolic AI approaches by identifying use cases that can benefit from these approaches.
- Invest in data architecture that can leverage the building blocks for neuro-symbolic AI techniques such as knowledge graphs and agent-based techniques.

Sample Vendors

Google Deepmind; Elemental Cognition; IBM; Microsoft; RelationalAI; Wolfram|Alpha

Gartner Recommended Reading

Innovation Insight: AI Simulation

Innovation Insight for Composite AI

Predicts 2023: Simulation Combined With Advanced AI Techniques Will Drive Future AI Investments

AI Engineering

Analysis By: Kevin Gabbard, Soyeb Barot

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

AI engineering is foundational for enterprise delivery of AI solutions at scale. The discipline unifies DataOps, MLOps and DevOps pipelines to create coherent enterprise development, delivery (hybrid, multicloud, edge), and operational (streaming, batch) AI-based systems.

Why This Is Important

The potential value of AI has led to huge demand to rapidly launch market-ready AI solutions. This is a big engineering challenge. Most enterprises still struggle to move individual pilots to production, much less operate portfolios of AI solutions at scale. Establishing consistent AI pipelines enables enterprises to develop, deploy, adapt and maintain AI models (statistical, machine learning, generative, deep learning, graph, linguistic and rule-based) consistently, regardless of environment.

Business Impact

AI engineering enables organizations to establish and grow high-value portfolios of AI solutions. Most AI development is currently limited by significant operational bottlenecks. With AI engineering methods — DataOps, ModelOps and DevOps — it is possible to deploy models into production in a structured, repeatable factory-model framework to realize significant value.

Drivers

- Intense hype surrounding generative AI solutions is increasing the overall demand for AI-powered solutions.
- DataOps, ModelOps and DevOps provide best practices for moving artifacts through the AI development life cycle. Standardization across data and model pipelines accelerates the delivery of AI solutions.
- The elimination of traditional siloed approaches to data management and AI engineering reduces impedance mismatch across data ingestion, processing, model engineering and deployment, which inevitably drift once the AI models are in production.
- AI engineering enables discoverable, composable and reusable AI artifacts (data catalogs, code repositories, reference architectures, feature stores, model stores, etc.) across the enterprise context. These are essential for scaling AI enterprisewide.
- AI engineering makes it possible to orchestrate solutions across hybrid, multicloud, edge AI or IoT.
- Broader use of foundational platforms provides initial success at scaling the production of AI initiatives with existing data, analytics and governance frameworks.

Obstacles

- Sponsorship for foundational enterprisewide AI initiatives is unclear.
- AI engineering requires simultaneous development of pipelines across domains.
- AI engineering requires integrating full-featured solutions with select tools, including open-source technologies, to address enterprise architecture gaps with minimal functional overlap. These include gaps around ETL stores, feature stores, model stores, model monitoring, pipeline observability and governance.
- AI engineering requires cloud maturity and possible rearchitecting, or the ability to integrate data and AI model pipelines across deployment contexts. Potential complexity and management of analytical and AI workloads alongside costs may deter organizations that are in the initial phases of AI initiatives.
- Enterprises often seek “unicorn” experts to productize AI platforms. Few vendors provide AI engineering capabilities, making such skills hard to find. Enterprises often have to build and support these environments on their own.

User Recommendations

- Establish a leadership mandate for enterprisewide foundational AI initiatives.
- Maximize business value from ongoing AI initiatives by establishing AI engineering practices that streamline the data, model and implementation pipelines.
- Simplify data and analytics pipelines by identifying the capabilities required to operationalize end-to-end AI platforms and build AI-specific toolchains.
- Use point solutions sparingly and only to plug feature/capability gaps in fully featured DataOps, MLOps, ModelOps and PlatformOps tools.
- Develop AI model management and governance practices that align model performance, human behavior and delivery of business value. Make it easier for users to adopt AI models.
- Leverage cloud service provider environments as foundational to build AI engineering. At the same time, rationalize your data, analytics and AI portfolios as you migrate to the cloud.
- Upskill data engineering and platform engineering teams to adopt tools and processes that drive continuous integration/continuous development for AI artifacts.

Sample Vendors

Amazon Web Services; Dataiku; DataRobot; Domino Data Lab; Google; HPE; IBM; Iguazio; Microsoft

Gartner Recommended Reading

Top Strategic Technology Trends for 2022: AI Engineering

Demystifying XOps: DataOps, MLOps, ModelOps, AIOps and Platform Ops for AI

A CTO's Guide to Top Artificial Intelligence Engineering Practices

Cool Vendors in Enterprise AI Operationalization and Engineering

Cool Vendors in AI Core Technologies — Scaling AI in the Enterprise

AI Simulation

Analysis By: Leinar Ramos, Anthony Mullen, Pieter den Hamer, Jim Hare

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

AI simulation is the combined application of AI and simulation technologies to jointly develop AI agents and the simulated environments in which they can be trained, tested and sometimes deployed. It includes both the use of AI to make simulations more efficient and useful, and the use of a wide range of simulation models to develop more versatile and adaptive AI systems.

Why This Is Important

Increased complexity in decision making is driving demand for both AI and simulation. However, current AI faces challenges, as it is brittle to change and requires a lot of data. Conversely, realistic simulations can be expensive and difficult to build and run. To resolve these challenges, a growing approach is to combine AI and simulation: Simulation is used to make AI more robust and compensate for a lack of training data, and AI is used to make simulations more efficient and realistic.

Business Impact

AI simulation can bring:

- Increased AI value by broadening its use to cases where data is scarce, using simulation to generate synthetic data (for example, robotics and self-driving cars)
- Greater efficiency by leveraging AI to decrease the time and cost to create and use complex and realistic simulations
- Greater robustness and adaptability by using simulation to generate diverse scenarios to increase AI performance in uncertain environments
- Decreased technical debt by reusing simulation environments to train future AI models

Drivers

- **Limited availability of AI training data is increasing the need for synthetic data techniques, such as simulation.** Simulation is uniquely positioned among synthetic data alternatives in its ability to generate diverse datasets that are not constrained by a fixed “seed” dataset to generate synthetic data from.

- **Advances in capabilities are making simulation increasingly useful for AI.** Simulation capabilities have been rapidly improving, driven both by increased computing performance and more efficient techniques. This has made simulation environments a key part of the training pipeline of some of the most advanced real-world AI use cases, such as robotics and self-driving cars.

- **The growing complexity of decision making is increasing the interest in AI simulation.** Simulation is able to generate diverse “corner case” scenarios that do not appear frequently in real-world data, but that are still crucial to train and test AI to perform well on uncertain environments. As the complexity of the environments and decision making goes up, the ability to build AI systems that are robust becomes more important.

- **Increased technical debt in AI is driving the need for the reusable environments that simulation provides.** Current AI focuses on building short-lived AI models with limited reuse, accumulating technical debt. Organizations will increasingly deploy hundreds of AI models, which requires a shift in focus toward building persistent, reusable environments where many AI models can be trained, customized and validated. Simulation environments are ideal since they are reusable, scalable, and enable the training of many AI models at once.

- **The growing sophistication of simulation drives the use of AI to make it more efficient.** Modern simulations are resource-intensive. This is driving the use of AI to accelerate simulation, typically by employing AI models that can replace parts of the simulation without running resource-intensive step-by-step numerical computations.

Obstacles

- **Gap between simulation and reality:** Simulations can only emulate — not fully replicate — real-world systems. This gap will reduce as simulation capabilities improve, but it will remain a key factor. Given this gap, AI models trained in simulation might not have the same performance once they are deployed: differences in the simulation training dataset versus real-world data can impact models' accuracy.

- **Complexity of AI simulation pipelines:** The combination of AI and simulation techniques can result in more complex pipelines that are harder to test, validate, maintain and troubleshoot.

- **Limited readiness to adopt AI simulation:** A lack of awareness among AI practitioners about leveraging simulation capabilities can prevent organizations from implementing an AI simulation approach.

- **Fragmented vendor market**: The AI and simulation markets are fragmented, with few vendors offering combined AI simulation solutions, potentially slowing down the deployment of this capability.

User Recommendations

- Complement AI with simulation to optimize business decision making or to overcome a lack of real-world data by offering a simulated environment for synthetic data generation or reinforcement learning.

- Complement simulation with AI by applying deep learning to accelerate simulation and generative AI to augment simulation.

- Create synergies between AI and simulation teams, projects and solutions to enable a next generation of more adaptive solutions for ever-more complex use cases. Incrementally build a common foundation of more generalized and complementary models that are reused across different use cases, business circumstances and ecosystems.

- Prepare for the combined use of AI, simulation and other relevant techniques, such as graphs, natural language processing or geospatial analytics, by prioritizing vendors that offer platforms that integrate different AI techniques (composite AI), as well as simulation.

Sample Vendors

Altair; Ansys; Cosmo Tech; Epic Games; MathWorks; Microsoft; NVIDIA; Rockwell Automation; The AnyLogic Company; Unity

Gartner Recommended Reading

Innovation Insight: AI Simulation

Predicts 2023: Simulation Combined With Advanced AI Techniques Will Drive Future AI Investments

Cool Vendors in Simulation for AI

Artificial General Intelligence

Analysis By: Pieter den Hamer

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Artificial general intelligence (AGI) is the (currently hypothetical) intelligence of a machine that can accomplish any intellectual task that a human can perform. AGI is a trait attributed to future autonomous AI agents that can achieve goals in a wide range of real or virtual environments at least as effectively as humans can. AGI is also called "strong AI."

Why This Is Important

As AI becomes more sophisticated and powerful, with recent great advances in generative AI in particular, a growing group of people see AGI as no longer purely hypothetical. Improving our understanding of at least the concept of AGI is critical for steering and regulating AI's further evolution. It is also important to manage realistic expectations and to avoid prematurely anthropomorphizing AI. However, should AGI become real, its impact on the economy, (geo)politics, culture and society cannot be underestimated.

Business Impact

In the short term, organizations must know that hype about AGI exists today among many stakeholders, stoking fears and unrealistic expectations about current AI's true capabilities. This AGI anticipation is already accelerating the emergence of more AI regulations and affects people's trust and willingness to apply AI today. In the long term, AI continues to grow in power and, with or without AGI, will increasingly impact organizations, including the advent of machine customers and autonomous business.

Drivers

- Recent great advances in applications of generative AI and the use of foundation models and large language or multimodal models drive considerable hype about AGI. These advances have been enabled largely by the massive scaling of deep learning, as well as by the availability of huge amounts of data and compute power. To further evolve AI toward AGI, however, current AI will need to be complemented by other (partially new) approaches, such as knowledge graphs, multiagent systems, simulations, evolutionary algorithms, causal AI, composite AI and likely other innovations yet unknown.

- Vendors such as Google, IBM, NNAISENSE, OpenAI and Vicarious are actively researching the field of AGI.

- Humans' innate desire to set lofty goals is also a major driver for AGI. At one point in history, humans wanted to fly by mimicking bird flight. Today, airplane travel is a reality. The inquisitiveness of the human mind, taking inspiration from nature and from itself, is not going to fizzle out.

- People's tendency to anthropomorphize nonliving entities also applies to AI-powered machines. This has been fueled by the humanlike responses of ChatGPT and similar AI, as well as AI being able to pass several higher-level education exams. In addition, more complex AI systems display behavior that has not been explicitly programmed. Among other reasons, this results from the dynamic interactions between many system components. As a result, AI is increasingly attributed with humanlike characteristics, such as understanding. Although many philosophers, neuropsychologists and other scientists consider this attribution as going too far or being highly uncertain, it has created a sense that AGI is within reach or at least is getting closer. In turn, this has triggered massive media attention, several calls for regulation to manage the risks of AGI and a great appetite to invest in AI for economic, societal and geopolitical reasons.

Obstacles

- The current issues regarding unreliability, hallucinations, lack of transparency and lack of reasoning or logic capabilities in generative AI-powered chatbots (one possible direction toward AGI), are not easy to overcome with the intrinsically probabilistic approach of deep learning. More data or more compute power for ever bigger models are unlikely to resolve these issues. Better or curated training data, improved prompt interpretation and engineering or more domain-specific foundation models may help to improve reliability, but not sufficiently.

- There is little scientific consensus about what “intelligence” and related terminology like “understanding” actually mean, let alone how AGI should be exactly defined and interpreted. Flamboyant representations of AGI in science fiction create a disconnect from reality. Scientific understanding about human intelligence is still challenged by the enormous complexity of the human brain and mind. Several breakthrough discoveries are still needed before human intelligence is properly understood at last. This in turn is foundational to the “design” or at least validation of AGI, even when AGI will emerge in a nonhuman, nonbrainlike form. Moreover, once AGI is understood and designed, further technological innovations will likely be needed to actually implement AGI. For these reasons, strong AI is unlikely to emerge in the near future. This may be sooner if one would settle for a more narrow, watered-down version of AGI in which AI is able to perform not all but only a few tasks at the same level as humans. This would no longer really be AGI as defined here.

- If AGI materializes, it is likely to lead to the emergence of autonomous actors that, in time, will be attributed with full self-learning, agency, identity and perhaps even morality. This will open up a bevy of legal rights of AI and trigger profound ethical and even religious discussions. Moreover, the (anticipated) emergence of AGI and the risk of human life being negatively impacted by AGI, from job losses to a new, AI-triggered arms race and more, may lead to a serious backlash and possibly regulatory bans on the development of AGI.

- The anticipated possible emergence of AGI urges governments to take measures before its risks can no longer be mitigated. Regulations to ban or control AGI are likely to emerge in the near future.

User Recommendations

- Today, people may be either overly concerned about future AI replacing humanity or overly excited about current AI's capabilities and impact on business. Both cases will hamper a realistic and effective approach toward using AI today. To mitigate this risk, engage with stakeholders to address their concerns and create or maintain realistic expectations.
- Stay apprised of scientific and innovative breakthroughs that may indicate the possible emergence of AGI. Meanwhile, keep applying current AI to learn, reap its benefits and develop practices for its responsible use.
- Although AGI is not a reality now, current AI already poses significant risks regarding bias, reliability and other areas. Adopt emerging AI regulations and promote internal AI governance to manage current and emerging future risks of AI.

Sample Vendors

AGI Innovations; Google; IBM; Kimera Systems; Microsoft; New Sapience; NNAISENSE; OpenAI; Vicarious

Gartner Recommended Reading

The Future of AI: Reshaping Society

Innovation Insight for Generative AI

Innovation Insight: AI Simulation

Applying AI — Key Trends and Futures

Innovation Insight for Artificial Intelligence Foundation Models

Causal AI

Analysis By: Pieter den Hamer, Leinar Ramos, Ben Yan

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Causal artificial intelligence (AI) identifies and utilizes cause-and-effect relationships to go beyond correlation-based predictive models and toward AI systems that can prescribe actions more effectively and act more autonomously. It includes different techniques, such as causal graphs and simulation, that help uncover causal relationships to improve decision making.

Why This Is Important

AI's ultimate value comes from helping people take better actions. Machine learning (ML) makes predictions based on statistical relationships (correlations), regardless of whether these are causal. This approach is fine for prediction, but predicting an outcome is not the same as understanding what causes it and how to improve it. Causal AI is crucial when we need to be more prescriptive to determine the best actions to influence specific outcomes. Causal AI techniques help make AI more autonomous, explainable, robust and efficient.

Business Impact

Causal AI leads to:

- Greater decision augmentation and autonomy in AI systems by estimating intervention effects
- Greater efficiency by adding domain knowledge to bootstrap AI models with smaller datasets
- Better explainability by capturing easy-to-interpret cause-and-effect relationships
- More robustness and adaptabilty by leveraging causal relationships that remain valid in changing environments
- The ability to extract causal knowledge with less costly and time-consuming experiments
- Reduced bias in AI systems by making causal links more explicit

Drivers

- **Analytics demand is shifting from predictive (what is likely to happen) to more prescriptive (what should be done) capabilities.** Making accurate predictions will remain key, but a causal understanding of how to affect predicted outcomes will be increasingly important.
- **AI systems increasingly need to act autonomously to generate business value,** particularly for time-sensitive and complex use cases, where human intervention is not feasible. This autonomy will only be possible by AI understanding what impact actions will have and how to make effective interventions.
- **Limited data availability for certain use cases is pushing organizations toward more data-efficient techniques like causal AI.** Causal AI leverages human domain knowledge of cause-and-effect relationships to bootstrap AI models in small-data situations.
- **The growing complexity of use cases and environments where AI is applied requires more robust AI techniques.** Causal structure changes much more slowly than statistical correlations, making causal AI more robust and adaptable in fast-changing environments. The volatility of the last few years has exposed the brittleness of correlation-based AI models across industries. These models have struggled to adapt because they were trained under a very different context.
- **The need for greater AI trust and explainability is driving interest in models that are more intuitive to humans.** Causal AI techniques, such as causal graphs, make it possible to be explicit about causes and explain models in terms that humans understand.
- **The next step in AI requires causal AI.** Current deep learning models and, in particular, generative AI have limitations in terms of their reliability and ability to reason. A composite AI approach that complements generative AI with causal AI — in particular, causal knowledge graphs — offers a promising avenue to bring AI to a higher level.

Obstacles

- **Causality is not trivial.** Not every phenomenon is easy to model in terms of its causes and effects. Causality might be unknown, regardless of AI use.
- **The quality of a causal AI model depends on its causal assumptions and on the data used to build it.** This data is susceptible to bias and imbalance. Just because a model is causal doesn't mean that it will outperform correlation-based ones.
- **Causal AI requires technical and domain expertise to properly estimate causal effects.** Building causal AI models is often more difficult than building correlation-based predictive models, requiring active collaboration between domain experts and AI experts.
- **AI experts might be unaware of causality methods.** If AI experts are overly reliant on data-driven models like ML, organizations could get pushback when looking to implement causal AI.
- **The vendor landscape is nascent, and enterprise adoption is currently low.** Clearly, this represents a challenge when organizations are running initial causal AI pilots and identifying specific use cases where causal AI is most relevant.

User Recommendations

- Acknowledge the limitations of correlation-based AI and ML approaches, which focus on leveraging correlations and mostly ignore causality. These limitations also apply to most generative AI techniques, including foundation models such as GPT-4.
- Use causal AI when you require more augmentation and automation in decision intelligence — i.e., when AI is needed not only to generate predictions, but also to understand how to affect the predicted outcomes. Examples include customer retention programs, marketing campaign allocation and financial portfolio optimization, as well as smart robotics and autonomous systems.
- Select different causal AI techniques depending on the complexity of the specific use case. These include causal rules, causal graphs and Bayesian networks, simulation, and ML for causal learning.
- Educate your data science teams on causal AI. Explain the difference between causal and correlation-based AI, and cover the range of techniques available to incorporate causality.

Sample Vendors

Actable AI; causaLens; Causality Link; CML Insight; Geminos Software; IBM; Lucid.AI; Qualcomm; SCALNYX; Xplain Data

Gartner Recommended Reading

Innovation Insight: Causal AI

Innovation Insight for Composite AI

Innovation Insight for Decision Intelligence Platforms

Building a Digital Future: Autonomic Business Operations

Case Study: Causal AI to Maximize the Efficiency of Business Investments (HDFC Bank)

Decision Intelligence

Analysis By: Erick Brethenoux

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Decision intelligence (DI) is a practical discipline that advances decision making by explicitly understanding and engineering how decisions are made and how outcomes are evaluated, managed and improved via feedback.

Why This Is Important

The current hype around automated decision making and augmented intelligence, fueled by AI techniques in decision making (including generative AI), is pushing DI toward the Peak of Inflated Expectations. Recent crises have revealed the brittleness of business processes. Reengineering those processes to be resilient, adaptable and flexible will require the discipline brought by DI methods and techniques. A fast-emerging market (DI platforms) is starting to provide resilient solutions for decision makers.

Business Impact

Decision intelligence helps:

- **Reduce technical debt and increase visibility.** It improves the impact of business processes by materially enhancing the sustainability of organizations' decision models based on the power of their relevance and the quality of their transparency, making decisions more transparent and auditable.
- **Reduce the unpredictability of decision outcomes.** It does so by properly capturing and accounting for the uncertain factors in the business context and making decision models more resilient.

Drivers

- **A dynamic and complex business environment, with an increasingly unpredictable and uncertain pace of business.** Two forces are creating a new market around decision intelligence platforms (DIPs). The first is the combination of AI techniques such as natural language processing, knowledge graphs and machine learning. The second is the confluence of several technology clusters around composite AI, smart business processes, insight engines, decision management and advanced personalization platforms.

- **The need to curtail unstructured, ad hoc decisions that are siloed and disjointed.** Often uncoordinated, such decisions promote local optimizations at the expense of global efficiency. This phenomenon happens from both an IT and a business perspective.

- **Expanding collaboration between humans and machines.** This collaboration, supplemented by a lack of trust in technologies (such as AI), is increasingly replacing tasks and promoting uneasiness from a human perspective. DI practices promote transparency, interpretability, fairness, reliability and accountability of decision models critical for the adoption of business-differentiating techniques.

- **Tighter regulations that are making risk management more prevalent.** From privacy and ethical guidelines to new laws and government mandates, it is becoming difficult for organizations to fully understand the risk impacts of their decisions. DI enables an explicit representation of decision models, reducing this risk.

- **Uncertainty regarding decision consistency across the organization.** Lack of explicit representation of decisions prevents proper harmonization of collective decision outcomes. DI remedies this issue.

- **Emergence of software tools in the form of decision intelligence platforms.** DIPs will enable organizations to practically implement DI projects and strategies.

- **Generative AI.** The advent of generative AI is accelerating the research and adoption of composite AI models, which are the foundation of DIPs.

Obstacles

- **Fragmentation:** Decision-making silos have created data, competencies and technology clusters that are difficult to reconcile and that could slow down the implementation of decision models.
- **Subpar operational structure:** An inadequate organizational structure around advanced techniques, such as the lack of an AI center of excellence, could impair DI progress.
- **Lack of proper coordination between business units:** The inability to impartially reconsider critical decision flows within and across departments (also because of fragmentation) diminishes the effectiveness of early DI efforts.
- **Lack of modeling in a wider context:** In organizations that have focused almost exclusively on technical skills, the other critical parts of human decision making — psychological, social, economic and organizational factors — have gone unaddressed.
- **Lack of AI literacy:** Many organizations still suffer from a lack of understanding when it comes to AI techniques. This AI illiteracy could slow down the development of DI projects.

User Recommendations

- **Promote the resiliency and sustainability of cross-organizational decisions** by building models using principles aimed at enhancing traceability, replicability, pertinence and trustworthiness.
- **Improve the predictability and alignment of decision agents** by simulating their collective behavior while also estimating their global contribution versus local optimization.
- **Develop staff expertise** in traditional and emerging decision augmentation and decision automation techniques, including predictive and prescriptive (optimization, business rules) analytics. Upskill business analysts, and develop new roles, such as decision engineer and decision steward.
- **Tailor the choice of decision-making technique** to the particular requirements of each decision situation by collaborating with subject matter experts, AI experts and business process analysts.
- **Accelerate the development of DI projects** by encouraging experimentation with generative AI and expediting the deployment of composite AI solutions.

Gartner Recommended Reading

Innovation Insight for Decision Intelligence Platforms

Reengineer Your Decision-Making Processes for More Relevant, Transparent and Resilient Outcomes

How to Choose Your Best-Fit Decision Management Suite Vendor

AI Security: How to Make AI Trustworthy

Top Strategic Technology Trends for 2023: Adaptive AI

Composite AI

Analysis By: Erick Brethenoux, Pieter den Hamer

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Composite AI refers to the combined application (or fusion) of different AI techniques to improve the efficiency of learning to broaden the level of knowledge representations. Composite AI broadens AI abstraction mechanisms and, ultimately, provides a platform to solve a wider range of business problems in a more effective manner.

Why This Is Important

Composite AI recognizes that no single AI technique is a silver bullet. Composite AI currently aims to combine “connectionist” AI approaches, like machine learning (ML), with “symbolic” and other AI approaches, like rule-based reasoning, graph analysis or optimization techniques. The goal is to enable AI solutions that require less data and energy to learn, embodying more abstraction mechanisms. Composite AI is at the center of the generative AI and decision intelligence market emergence.

Business Impact

Composite AI offers two main benefits. First, it brings the power of AI to a broader group of organizations that do not have access to large amounts of historical or labeled data but possess significant human expertise. Second, it helps to expand the scope and quality of AI applications (that is, more types of reasoning challenges can be embedded). Other benefits, depending on the techniques applied, include better interpretability and resilience and the support of augmented intelligence.

Drivers

- **ML-based AI techniques lead to insights that inform actions.** Additionally, the most appropriate actions can be further determined by combinations of rule-based and optimization models — a combination often referred to as prescriptive analytics.
- **Small datasets, or the limited availability of data, have pushed organizations to combine multiple AI techniques.** Where raw historical data has been more scarce, enterprises have started to complement it using additional AI techniques, such as knowledge graphs and generative adversarial networks (GANs), to generate synthetic data.
- **Combining AI techniques is much more effective than relying only on heuristics or a fully data-driven approach.** A heuristic or rule approach can be combined with a deep learning model (for example, predictive maintenance). Rules coming from human experts, or the application of physical/engineering model analysis, may specify that certain sensor readings indicate inefficient asset operations. This can be used as a feature to train a neural network to assess and predict the asset's health, also integrating causal AI capabilities.
- **Proliferation of computer vision and NLP solutions.** In computer vision, (deep) neural networks are used to identify or categorize people or objects in an image. This output can be used to enrich or generate a graph, representing the image entities and their relationships.
- **Agent-based modeling is the next wave of composite AI.** A composite AI solution can be composed of multiple agents, each representing an actor in the ecosystem. Combining these agents into a "swarm" enables the creation of common situation awareness, more global planning optimization, responsive scheduling and process resilience.
- **The acceleration of generative AI.** The advent of generative AI is accelerating the research and adoption of composite AI models (through artifacts, process and collaboration generations), which are the foundation of decision intelligence (DI) platforms.

Obstacles

- **Lack of awareness and skills in leveraging multiple AI methods.** This could prevent organizations from considering the techniques particularly suited to solving specific problem types.
- **Deploying ModelOps.** The ModelOps domain (i.e., the operationalization of multiple AI models, such as optimization models, rule models and graph models) remains an art much more than a science. A robust ModelOps approach will be necessary to efficiently govern composite AI environments and harmonize it with other disciplines, such as DevOps and DataOps.
- **Trust and risk barriers.** The AI engineering discipline is also starting to take shape, but only mature organizations have started to apply its benefits in operationalizing AI techniques. Security, ethical model behaviors, observability, model autonomy and change management practices will have to be addressed across the combined AI techniques.

User Recommendations

- **Identify projects in which a fully data-driven, ML-only approach is inefficient or ill-fitted.** For example, in cases when enough data is not available or when the pattern cannot be represented through current ML models.
- **Capture domain knowledge and human expertise** to provide context for data-driven insights by applying decision management with business rules and knowledge graphs, in conjunction with ML and/or causal models.
- **Combine the power of ML, image recognition or natural language processing with graph analytics** to add higher-level, symbolic and relational intelligence.
- **Extend the skills of ML experts, or recruit/upskill additional AI experts,** to also cover graph analytics, optimization or other techniques for composite AI. For rules and heuristics, consider knowledge engineering skills, as well as emerging skills such as prompt engineering.
- **Accelerate the development of DI projects** by encouraging experimentation with generative AI, which will in turn accelerate the deployment of composite AI solutions.

Sample Vendors

ACTICO; Aera Technology; FICO; Frontline Systems; IBM; Indico Data; Peak; SAS

Gartner Recommended Reading

How to Use Machine Learning, Business Rules and Optimization in Decision Management

Top Strategic Technology Trends for 2022: AI Engineering

Innovation Insight for Decision Intelligence

Innovation Insight for Decision Intelligence Platforms

How to Choose Your Best-Fit Decision Management Suite Vendor

Data-Centric AI

Analysis By: Svetlana Sicular

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Embryonic

Definition:

Data-centric AI is an approach that focuses on enhancing and enriching training data to drive better AI outcomes, as opposed to a model-centric approach wherein AI outcomes are driven by model tuning. Data-centric AI also addresses data quality, privacy and scalability. Quality data is crucial for generative AI to perform well on specific tasks. Examples of data-centric AI include knowledge graphs, feature stores, synthetic data, data labeling and annotation, and federated ML.

Why This Is Important

The AI community is facing a bifurcation of data-centric and model-centric AI, because data quality, curation and consistency often improve AI accuracy more efficiently than tweaking models. Data-centric AI disrupts classical data management and prevalent model-centric data science by addressing AI-specific data considerations to improve the quality of models on an ongoing basis. Data-centric AI is becoming especially important with the rise of pretrained off-the-shelf models.

Business Impact

Organizations that invest in AI at scale will shake up their data management practices and capabilities to preserve the evergreen classical ideas and extend them to AI in two ways:

- Add capabilities necessary for convenient AI development by an AI-focused audience that is not familiar with data management, such as bias mitigation and data drift monitoring.
- Use AI to improve and augment evergreen classics of data governance, persistence, integration and data quality.

Drivers

- **Models, especially for generative AI, increasingly come from the vendors, rather than being delivered in-house.** Data is becoming the main means for enterprises to get value from these pretrained models.
- **Most commonly delivered AI solutions depend on data availability, quality and understanding, not just AI model building.** Many enterprises attempt to tackle AI without considering AI-specific data management issues. The importance of data management in AI is often underestimated, so data management solutions are now being adjusted for AI needs.
- **Classical data management is ripe for disruption, to support AI efforts.** Rapid progress of AI poses new challenges in organizing and managing the data for AI. We expect a cycle of augmented data management techniques that are better suited for meeting the data requirements of AI. Data ecosystems on the foundation of data fabric indicate the beginning of this new cycle.
- **Data management capabilities, practices and tools greatly benefit AI development and deployment.**The AI community invents new data-centric approaches and it also opens up to data management innovations like data fabric and lakehouse.
- **New data management solutions mitigate AI-amplified bias** originating in data interpretation, labeling and human actions recorded in the data. Bias mitigation is an acute, AI-specific problem. AI bias is the use case in data management calling to determine how to structure, analyze and prepare data.

Obstacles

- **Data-centric AI is disconnected from data management.** The AI community remains oblivious to data management capabilities, practices and tools that can greatly benefit AI development and deployment. Traditional data management also ignores the AI-specific considerations, such as data bias, labeling and drift; this is changing but slowly.
- **Even though the data side of AI reflects understanding of the problem, it is less exciting.** It includes tasks such as preparing datasets and developing a clear understanding of why the data was collected a certain way, what the data means and what biases exist in the data.
- **Responsible AI is necessary to ask all the right questions about the data and the solution.** These are AI-specific data practices that many enterprises want to solve through tooling, rather than governance.
- **Data management activities don’t end once the model has been developed.** Deployment considerations and ongoing drift monitoring require dedicated data management activities and practices.

User Recommendations

- **Formalize data-centric AI and model-centric data as part of your data management strategy.** Implement active metadata and data fabric as foundational components of this strategy. Combine classical and new capabilities to meet AI needs. Establish roles and responsibilities to manage data in support of AI.
- **Approach model development in a data-centric way** due to the dependency of AI models on quality data. Diversify data, models and people to ensure AI value and ethics.
- **Leverage data management expertise, AI engineering, DataOps and MLOps approaches** to support the AI life cycle. Include data management requirements when deploying models. Develop data monitoring and data governance metrics to ensure that your AI models produce the correct output.
- **Enforce policies on data fitness for AI.** Define and measure minimum data standards (such as formats, tools and metrics) for AI early on, to prevent reconciliation of multiple data approaches when taking AI to scale.

Sample Vendors

Databricks; Explorium; Landing AI; Mobilewalla; MOSTLY AI; Pinecone Systems; Protopia AI; Scale AI; Snorkel AI; YData

Gartner Recommended Reading

Cool Vendors in Data-Centric AI

Data Science and Machine Learning Trends You Can't Ignore

Securing Your AI Data Pipeline

Overcoming Data Quality Risks When Using Semistructured and Unstructured Data for AI/ML Models

5 Ways to Enhance Your Data Engineering Practices

Operational AI Systems

Analysis By: Chirag Dekate, Soyeb Barot, Sumit Agarwal

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Operational AI systems (OAISys) enable orchestration, automation and scaling of production-ready and enterprise-grade AI, comprising ML, DNNs and Generative AI. OAISys integrates DataOps, ModelOps, MLOps and deployment services to deliver enterprise-grade governance, including reusability, reproducibility, release management, lineage, risk and compliance management, and security. It also unifies development, delivery (hybrid, multicloud, IoT) and operational (streaming, batch) contexts.

Why This Is Important

OAISys can help enterprises:

- Standardize, govern and automate AI engineering and deployment technologies, and accelerate productization of AI.

- Eliminate system integration friction and impedance mismatch across DataOps, ModelOps, MLOps, deployment and governance platforms.
- Scale AI initiatives by enabling orchestration across hybrid, multicloud, edge AI or IoT.
- Enable discoverable, composable and reusable AI artifacts (data catalogs, feature stores, model stores) across the enterprise context.

Business Impact

OAISys deliver production AI systems that:

- Systemize analytics and AI engineering technologies, including ModelOps and MLOps platforms.
- Integrate existing data, analytics and DSML platforms.
- Utilize reusability components including feature and model stores, monitoring, experiment management, model performance and lineage tracking.
- Homogenize governance including compliance, risk, security, and cost across deployment (hybrid, multicloud, IoT) and operational (streaming, batch) contexts.

Drivers

- Enable business stakeholders to leverage AI as a service that is customized to their enterprise context.
- IT leaders need to deliver, manage and govern AI models within enterprise applications deployed across multiple contexts and jurisdictions (hybrid, multicloud, edge AI and IoT).
- Traditional siloed approaches of data management and AI engineering create integration challenges across the data ingest, processing, model engineering and deployment.
- OAISys enables enterprises to standardize and automate development, management, deployment, maintenance and governance technologies to deliver comprehensive, flexible and composed end-to-end AI systems.
- It helps align and automate the data, AI model deployment and governance pipelines.
- Operationalization and automation platforms are a core part of how early enterprise AI pioneers scale productization of AI by leveraging existing data, analytics and governance frameworks.
- Standardizing data pipelines, including DataOps toolchains, creating reusability components such as data catalogs and ETL registries, monitoring, security, access control and lineage tracking.
- The enterprise OAISys enables unification of two core contexts: deployment context across hybrid, multicloud, edge AI and IoT, and operational context across batch and streaming processing modes that commonly occur as enterprises train and deploy production models.

Obstacles

- Enterprises with low data and AI maturity levels will find OAISys intimidating to build, deliver and support.
- OAISys requires integration of full-featured solutions with select tools that address portfolio gaps with minimal overlap. These include capability gaps around feature stores, model stores, governance capabilities and more.
- OAISys requires a high degree of cloud maturity, or the ability to integrate data and model pipelines across deployment contexts. The potential complexity and costs may be a deterrent for organizations just starting their AI initiatives.
- Enterprises seeking to deliver OAISys often seek "unicorn" experts and service providers to productize AI. Fully featured vendor solutions that enable OAISys are hard to come by, and enterprises often have to build and support these environments on their own.

User Recommendations

- Focus AI engineering activities to deliver business context customized operational AI systems.
- Rationalize data and analytic environment and leverage current (simplified subset of) investments in data management, DSML, ModelOps and MLOps tools to build OAISys.
- Leverage cloud service provider environments as foundational environments to build OAISys along with rationalizing your data, analytics and AI portfolios as you migrate to the cloud.
- Avoid building patchwork OAISys that integrate piecemeal functionality from scratch (and add another layer of tool sprawl). Utilize point solutions sparingly and surgically to plug feature/capability gaps in fully featured DataOps, MLOps and ModelOps tools.
- Actively leverage your existing data management, DSML, MLOps and ModelOps platforms as building blocks, rather than starting from scratch.

Sample Vendors

Amazon Web Services; Dataiku; DataRobot; Domino Data Lab; Google; HPE Ezmeral Software; IBM; Iguazio; Microsoft; ModelOp

Gartner Recommended Reading

2023 Planning Guide for Analytics and Artificial Intelligence

Emerging Tech Impact Radar: Data and Analytics

Quick Answer: How Should CXOs Structure AI Operating Models?

At the Peak

AI TRiSM

Analysis By: Avivah Litan, Jeremy D'Hoinne, Bart Willemsen

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

AI trust, risk and security management (AI TRiSM) ensures AI model governance, trustworthiness, fairness, reliability, robustness, efficacy and data protection. AI TRiSM includes solutions and techniques for model interpretability and explainability, data and content anomaly detection, AI data protection, model operations and adversarial attack resistance.

Why This Is Important

AI models and applications deployed in production should be subject to protection mechanisms. Doing so ensures sustained value generation and acceptable use based on predetermined intentions. Accordingly, AI TRiSM is a framework that comprises a set of risk and security controls and trust enablers that helps enterprises govern and manage AI models and applications' life cycle — and accomplish business goals. The collateral benefit is enhanced compliance with forthcoming regulations, like the EU AI Act.

Business Impact

Organizations that do not consistently manage AI risks are exponentially inclined to experience adverse outcomes, such as project failures and breaches. Inaccurate, unethical or unintended AI outcomes, process errors and interference from malicious actors can result in security failures, financial and reputational loss or liability, and social harm. AI misperformance can also lead organizations to make suboptimal business decisions.

Drivers

- ChatGPT democratized third-party-provisioned generative AI and transformed how enterprises compete and do work. Accordingly, the risks associated with hosted, cloud-based generative AI applications are significant and rapidly evolving.

- Democratized, third-party-provisioned AI often poses considerable data confidentiality risks. This is because large, sensitive datasets used to train AI models are shared across organizations. Confidential data access must be carefully controlled to avoid adverse regulatory, commercial and reputational consequences.
- AI risk and security management imposes new operational requirements that are not fully understood and cannot be addressed by existing systems. New vendors are filling this gap.
- AI models and applications must be constantly monitored to ensure that implementations are compliant, fair and ethical. Risk management tools can identify and eliminate bias from training data and AI algorithms.
- AI model explainability must be constantly tested through model observations. Doing so ensures original explanations and interpretations of AI models remain active during model operations. If they don't, corrective actions must be taken.
- Detecting and stopping adversarial attacks on AI requires new methods that most enterprise security systems do not offer.
- Regulations for AI risk management — such as the EU AI Act and other regulatory frameworks in North America, China and India — are driving businesses to institute measures for managing AI model application risk. Such regulations define new compliance requirements organizations will have to meet on top of existing ones, like those pertaining to privacy protection.

Obstacles

- AI TRiSM is often an afterthought. Organizations generally don't consider it until models or applications are in production.
- Enterprises interfacing with hosted, large language models (LLMs) are missing native capabilities to automatically filter inputs and outputs — for example, confidential data policy violations or inaccurate information used for decision making. Also, enterprises must rely on vendor licensing agreements to ensure their confidential data remains private in the host environment.
- Once models and applications are in production, AI TRiSM becomes more challenging to retrofit to the AI workflow, thus creating inefficiencies and opening the process to potential risks.
- Most AI threats are not fully understood and not effectively addressed.

- AI TRiSM requires a cross-functional team, including legal, compliance, security, IT and data analytics staff, to establish common goals and use common frameworks — which is difficult to achieve.
- Although challenging, the integration of life cycle controls can be done with AI TRiSM.

User Recommendations

- Set up an organizational task force or dedicated unit to manage your AI TRiSM efforts. Include members who have a vested interest in your organization's AI projects.
- Work across your organization to effectively manage best-of-breed toolsets for enterprise-managed AI and applications that use hosted AI as part of a comprehensive AI TRiSM program.
- Avoid, to the extent possible, black-box models that stakeholders do not understand.
- Implement solutions that protect data used by AI models. Prepare to use different methods for different use cases and components.
- Establish data protection and privacy assurances in license agreements with vendors hosting LLM models — for example, Microsoft or OpenAI.
- Use enterprise-policy-driven content filtering for inputs and outputs to and from hosted models, such as LLMs.
- Incorporate risk management mechanisms into AI models and applications' design and operations. Constantly validate reliable and acceptable use cases.

Sample Vendors

AIShield; Arize AI; Arthur; Fiddler; ModelOp; Modzy; MOSTLY AI; Protopia AI; SolasAI; TrojAI

Gartner Recommended Reading

Use Gartner's MOST Framework for AI Trust and Risk Management

Top 5 Priorities for Managing AI Risk Within Gartner's MOST Framework

Prompt Engineering

Analysis By: Frances Karamouzis, Afraz Jaffri, Jim Hare, Arun Chandrasekaran, Van Baker

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Prompt engineering is the discipline of providing inputs, in the form of text or images, to generative AI models to specify and confine the set of responses the model can produce. The inputs prompt a set that produces a desired outcome without updating the actual weights of the model (as done with fine-tuning). Prompt engineering is also referred to as “in-context learning,” where examples are provided to further guide the model.

Why This Is Important

Prompt engineering is the linchpin to business alignment for desired outcomes. Prompt engineering is important because large language models (LLMs) and generative AI models in general are extremely sensitive to nuances and small variations in input. A slight tweak can change an incorrect answer to one that is usable as an output. Each model has its own sensitivity level, and the discipline of prompt engineering is to uncover the sensitivity through iterative testing and evaluation.

Business Impact

Prompt engineering has the following business impacts:

- **Performance:** It helps improve model performance and reduce hallucinations.
- **Business alignment**: It allows subject data scientists, subject matter experts and software engineers to steer foundation models, which are general-purpose in nature, to align to the business, domain and industry.
- **Efficiency and effectiveness:** Alternative options, such as building a model from scratch or fine-tuning, can be much more complex, drive longer time to market and be more expensive.

Drivers

- **Balance and efficiency:** The fundamental driver for prompt engineering is it allows organizations to strike a balance between consuming an "as is" offering versus pursuing a more expensive and time-consuming approach of fine-tuning. Generative AI models, and in particular LLMs, are pretrained, so the data that enterprises want to use with these models cannot be added to the training set. Instead, prompts can be used to feed content to the model with an instruction to carry out a function.

- **Process or task-specific customizations or new use cases:** The insertion of context and patterns that a model uses to influence the output generated allows for customizations for a particular enterprise or domain, or regulatory items. Prompts are created to help improve the quality for different use cases — such as domain-specific question answering, summarization, categorization, and so on — with or without the need for fine-tuning a model, which can be expensive or impractical. This would also apply to creating and designing new use cases that utilize the model's capability for image and text generation.

- **Validation and verification:** It is important to test, understand and document the limits and weaknesses of the models to ensure a reduced risk of hallucination and unwanted outputs.

Obstacles

- **Embryonic nature of the discipline:** Prompt engineering processes and roles are either unknown or enterprises have a low level of understanding and experience. Gartner webinar polling data (over 2,500 responses; see Executive Pulse: AI Investment Gets a Boost From ChatGPT Hype) revealed that approximately 60% of respondents self-reported that they had not heard of prompt engineering. And 90% of those same respondents revealed that their organization did not currently have prompt engineers.

- **Role alignment:** Data scientists are critical to understanding the capabilities and limits of models, and to determine whether to pursue a purely prompt-based or fine-tuning-based approach (or combination of approaches) for customization. The ultimate goal is to use machine learning itself to generate the best prompts and achieve automated prompt optimization. This is in contrast to an end user of an LLM who concentrates on prompt design to manually alter prompts to give better responses.

- **Lack of business alignment:** There is often a lack of consensus on prompt engineering's business approach, as well as agreed-upon standards, methodology and approaches. This has led to fierce debates on the value of prompt engineering and how to establish governance.

- **Risk:** Beyond the early stages of awareness and understanding, the biggest obstacle may be that prompt engineering is focused on verification, validation, improvement and refinement; however, it's not without risk. Prompt engineering is not the panacea to all of the challenges. It helps to manage risk, not remove it completely. Errors may still occur, and potential liability is at stake.

User Recommendations

- Rapidly build awareness and understanding of prompt engineering in order to quickly start the journey of shape-shifting the appropriate prompt engineering discipline and teams.
- Build critical skills across a number of different team members that will synergistically contribute critical elements. For example, there are important roles for data scientists, business users, domain experts, software engineers and citizen developers.
- Communicate and cascade the message that prompt engineering is not foolproof. Rigor and diligence need to permeate and work across all the enterprise teams to ensure successful solutions.

Sample Vendors

FlowGPT; HoneyHive; LangChain; PromptBase; Prompt Flow; PromptLayer

Gartner Recommended Reading

Quick Answer: How Will Prompt Engineering Impact the Work of Data Scientists?

Quick Answer: What Impact Will Generative AI Have on Search?

Accelerate Adoption of Generative AI by Offering an FMOps- or a Domain-Specific Partner Ecosystem

Glossary of Terms for Generative AI and Large Language Models

Neuromorphic Computing

Analysis By: Alan Priestley

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Neuromorphic computing is a technology that provides a mechanism to more accurately model the operation of a biological brain using digital or analog processing techniques. These designs typically use spiking neural networks (SNNs), rather than the deep neural networks (DNNs) of the current generations of AI technologies, feature non-von Neumann architectures and are characterized by simple processing elements, but very high interconnectivity.

Why This Is Important

Currently, most AI development leverages parallel processing designs based on GPUs. These are high-performance, but high-power-consuming, devices that are not applicable in many deployments. Neuromorphic computing utilizes asynchronous, event-based designs that have the potential to offer extremely low power operation. This makes them uniquely suitable for edge and endpoint devices, where their ability to support object and pattern recognition can enable image, audio and sensor analytics.

Business Impact

AI techniques are rapidly evolving, enabled by radically new computing designs.

- Today’s deep neural network (DNN) algorithms require the use of high-performance processing devices and vast amounts of data to train these systems, limiting scope of deployment.
- Neuromorphic computing designs can be implemented using low-power devices, bringing the potential to drive the reach of AI techniques out to the edge of the network, accelerating key tasks such as image and sound recognition.

Drivers

- Different design approaches are being taken to implement neuromorphic computing designs — large-scale devices for use in data centers, and smaller-scale devices for edge computing and endpoint designs. Both these paths leverage spiking neural networks (SNNs) to implement asynchronous designs that have the benefit of being extremely low power when compared with current DNN-based designs.

- Semiconductor vendors are developing chips that utilize SNNs to implement AI-based solutions.

- Neuromorphic computing architectures have the potential to deliver extreme performance for use cases such as DNNs and signal analysis at very low power.

- Neuromorphic systems can be trained using smaller datasets than DNNs, with the potential of in situ training.

Obstacles

- Accessibility: GPUs are more accessible and easier to program than neuromorphic computing. However, this could change when neuromorphic computing and the supporting ecosystems mature.

- Knowledge gaps: Programming neuromorphic computing will require new programming models, tools and training methodologies.

- Scalability: The complexity of interconnection challenges the ability of semiconductor manufacturers to create viable neuromorphic devices.

- Integration: Significant advances in architecture and implementation are required to compete with other DNN-based architectures. Rapid developments in DNN architectures may slow advances in neuromorphic computing, but there are likely to be major leaps forward in the next decade.

User Recommendations

- Prepare for future utilization as neuromorphic architectures have the potential to become viable over the next five years.
- Create a roadmap plan by identifying key applications that could benefit from neuromorphic computing.
- Partner with key industry leaders in neuromorphic computing to develop proof-of-concept projects.
- Identify new skill sets required to be nurtured for successful development of neuromorphic initiatives, and establish a set of business outcomes/expected value to set management's long-term expectations.

Sample Vendors

AnotherBrain; Applied Brain Research; BrainChip; GrAi Matter Labs; Intel; Natural Intelligence; SynSense

Gartner Recommended Reading

Emerging Technologies: Tech Innovators in Neuromorphic Computing

Emerging Technologies: Top Use Cases for Neuromorphic Computing

Forecast: AI Semiconductors, Worldwide, 2021-2027

Emerging Tech Impact Radar: Artificial Intelligence

Responsible AI

Analysis By: Svetlana Sicular

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Responsible artificial intelligence (AI) is an umbrella term for aspects of making appropriate business and ethical choices when adopting AI. These include business and societal value, risk, trust, transparency, fairness, bias mitigation, explainability, sustainability, accountability, safety, privacy, and regulatory compliance. Responsible AI encompasses organizational responsibilities and practices that ensure positive, accountable, and ethical AI development and operation.

Why This Is Important

Responsible AI has emerged as the key AI topic for Gartner clients. When AI replaces human decisions and generates brand-new artifacts, it amplifies both good and bad outcomes. Responsible AI enables the right outcomes by ensuring business value while mitigating risks. This requires a set of tools and approaches, including industry-specific methods, adopted by vendors and enterprises. More jurisdictions introduce new regulations that challenge organizations to respond in meaningful ways.

Business Impact

Responsible AI assumes accountability for AI development and use at the individual, organizational and societal levels. If AI governance is practiced by designated groups, responsible AI applies to everyone involved in the AI process. Responsible AI helps achieve fairness, even though biases are baked into the data; gain trust, although transparency and explainability methods are evolving; and ensure regulatory compliance, despite the AI's probabilistic nature.

Drivers

- Responsible AI means a deliberate approach in many directions at once. Data science's responsibility to deliver unbiased, trusted and ethical AI is just the tip of the iceberg. Responsible AI helps AI participants develop, implement, utilize and address the various drivers they face.
- Organizational driver assumes that AI's business value versus risk in regulatory, business and ethical constraints should be balanced, including employee reskilling and intellectual property protection.
- Societal driver includes resolving AI safety for societal well-being versus limiting human freedoms. Existing and pending legal guidelines and regulations, such as the EU's Artificial Intelligence Act, make responsible AI a necessity.
- Customer/citizen driver is based on fairness and ethics and requires resolving privacy versus convenience. Customers should exhibit readiness to give their data in exchange for benefits. Consumer and citizen protection regulations provide the necessary steps, but do not relieve organizations of deliberation specific to their constituents.
- With further AI adoption, the responsible AI framework is becoming more important and is better understood by vendors, buyers, society and legislators.
- AI affects all ways of life and touches all societal strata; hence, the responsible AI challenges are multifaceted and cannot be easily generalized. New problems constantly arise with rapidly evolving technologies and their uses, such as using OpenAI's ChatGPT or detecting deepfakes. Most organizations combine some of the drivers under the umbrella of responsible AI, namely, accountability, diversity, ethics, explainability, fairness, human centricity, operational responsibility, privacy, regulatory compliance, risk management, safety, transparency and trustworthiness.

Obstacles

- Poorly defined accountability for responsible AI makes it look good on paper but is ineffective in reality.
- Unawareness of AI's unintended consequences persists. Forty percent of organizations had an AI privacy breach or security incident. Many organizations turn to responsible AI only after they experience AI's negative effects, whereas prevention is easier and less stressful.
- Legislative challenges lead to efforts for regulatory compliance, while most AI regulations are still in draft. AI products' adoption of regulations for privacy and intellectual property makes it challenging for organizations to ensure compliance and avoid all possible liability risks.
- Rapidly evolving AI technologies, including tools for explainability, bias detection, privacy protection and some regulatory compliance, lull organizations into a false sense of responsibility, while mere technology is not enough. A disciplined AI ethics and governance approach is necessary, in addition to technology.

User Recommendations

- Publicize consistent approaches across all focus areas. The most typical areas of responsible AI in the enterprise are fairness, bias mitigation, ethics, risk management, privacy, sustainability and regulatory compliance.
- Designate a champion accountable for the responsible development and use of AI for each use case.
- Define model design and exploitation principles. Address responsible AI in all phases of model development and implementation cycles. Go for hard trade-off questions. Provide responsible AI training to personnel.
- Establish operationalize responsible AI principles. Ensure diversity of participants and the ease to voice AI concerns.
- Participate in industry or societal AI groups. Learn best practices and contribute your own, because everybody will benefit from this. Ensure policies account for the needs of any internal or external stakeholders.

Sample Vendors

Amazon; Arthur; Fiddler; Google; H2O.ai; IBM; Microsoft; Responsible AI Institute; TAZI.AI; TruEra

Gartner Recommended Reading

A Comprehensive Guide to Responsible AI

Expert Insight Video: What Is Responsible AI and Why Should You Care About It?

Best Practices for the Responsible Use of Natural Language Technologies

Activate Responsible AI Principles Using Human-Centered Design Techniques

How to Ensure Your Vendors Are Accountable for Governance of Responsible AI

Smart Robots

Analysis By: Annette Jump

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

A smart robot is an AI-powered, often-mobile machine designed to autonomously execute one or more physical tasks. These tasks may rely on, or generate, machine learning, which can be incorporated into future activities or support unprecedented conditions. Smart robots can be split into different types based on the tasks/use cases, such as personal, logistics and industrial.

Why This Is Important

Smart robotics is an AI use case, while robotics in general does not imply AI. Smart (physical) robots had less adoption compared with industrial counterparts but received great hype in the marketplace; therefore, smart robots are still climbing the Peak of Inflated Expectations. There has been an increased interest in smart robots in the last 12 months, as companies are looking to further improve logistic operations, support automation and augment humans in various jobs.

Business Impact

Smart robots will make their initial business impact across a wide spectrum of asset-, product- and service-centric industries. Their ability to reduce physical risk to humans, as well as do work with greater reliability, lower costs and higher productivity, is common across these industries. Smart robots are already being deployed among humans to work in logistics, warehousing, police as well as safety applications.

Drivers

- The market is becoming more dynamic with technical developments of the last two years, enabling a host of new use cases that have changed how smart robots are perceived and how they can deliver value.
- The physical building blocks of smart robots (motors, actuators, chassis and wheels) have incrementally improved over time. However, areas such as Internet of Things (IoT) integration, edge AI and conversational capabilities have seen fundamental breakthroughs. This changes the paradigm for robot deployments.
- Vendor specialization has increased, leading to solutions that have higher business value, since an all-purpose/multipurpose device is either not possible or is less valuable.
- Growing interest in smart robots across a broad number of industries and use cases like: medical/healthcare (patient care, medical materials handling, interdepartment deliveries and sanitization); manufacturing (product assembly, stock replenishment, support of remote operations and quality control [QC] check); last-mile delivery; inspection of industrial objects or equipment; agriculture (harvesting and processing crops); and workplace and concierge robots in workplaces, hospitality, hospitals and so forth.

Obstacles

- **Companies are still struggling to identify valuable business use cases and assess ROI** for robots, especially outside of manufacturing and transportation. Therefore, the position of "smart robots" is still climbing to the Peak of Inflated Expectations.
- **Hype and expectations will continue to build around smart robots during the next few years**, as providers expand their offerings and explore new technologies, like reinforcement learning to drive a continuous loop of learning for robots and swarm management.
- **Lack of ubiquitous wireless connectivity solutions outside of smart spaces and immaturity of edge AI technologies** can inhibit the pace at which smart robots become semiautomated and mobile.
- **The need to offload computation to the cloud** will decrease from 2024, as robots will make more autonomous decisions.
- **The continuous evolution of pricing models**, like buy, monthly lease or hourly charge versus robot as a service for robotic solutions can create some uncertainty for organizations.

User Recommendations

- Evaluate smart robots as both substitutes and complements to their human workforce in manufacturing, distribution, logistics, retail, healthcare or defense.
- Begin pilots designed to assess product capability and quantify benefits, especially as ROI is possible even with small-scale deployments.
- Examine current business processes for current deployment of smart robots and also for large-scale deployment over the next three to five years.
- Consider different purchase models for smart robots.
- Dissolve the reluctance from staff by developing training resources to introduce robots alongside humans as an assistant.
- Ensure there are sufficient cloud computing resources to support high-speed and low-latency connectivity in the next two years.
- Evaluate multiple global and regional providers due to fragmentation within the robot landscape.

Sample Vendors

Ava Robotics; Geek+; GreyOrange; iRobot; Locus Robotics; Rethink Robotics; SoftBank Robotics; Symbotic; Temi; UBTECH

Gartner Recommended Reading

Emerging Technologies: Top Use Cases for Smart Robots to Lead the Way in Human Augmentation

Emerging Technologies: Top Use Cases Where Robots Interact Directly With Humans

Emerging Technologies: Venture Capital Growth Insights for Robots, 2021

Emerging Technologies: Smart Robot Adoption Generates Diverse Business Value

Foundation Models

Analysis By: Arun Chandrasekaran

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Foundation models are large-parameter models that are trained on a broad gamut of datasets in a self-supervised manner. They are mostly based on transformer or diffusion deep neural network architectures and will potentially be multimodal in the near future. They are called foundation models because of their critical importance and applicability to a wide variety of downstream use cases. This broad applicability is due to the pretraining and versatility of the models.

Why This Is Important

Foundation models are an important step forward for AI due to their massive pretraining and wide use-case applicability. They can deliver state-of-the-art capabilities with higher efficacy than their predecessors. They've become the go-to architecture for NLP, and have also been applied to computer vision, audio and video processing, software engineering, chemistry, finance, and legal use cases. Primarily text-based, large language models (LLMs) are a popular subset of foundation models. ChatGPT is based on one (GPT-4).

Business Impact

With their potential to enhance applications across a broad range of natural language use cases, foundation models will have a wide impact across vertical industries and business functions. Their impact has accelerated, with a growing ecosystem of startups building enterprise applications on top of them. Foundation models will advance digital transformation within the enterprise by improving workforce productivity, automating and enhancing CX, and enabling rapid, cost-effective creation of new products and services.

Drivers

Foundation models:

- **Require only limited model customization to deliver effective results.** Foundation models can effectively deliver value through prebuilt APIs, prompt engineering or further fine-tuning. While fine-tuning may deliver the best value because of customization possibilities, the other two options are less complex.

- **Deliver superior natural language processing.** The difference between these models and prior neural network solutions is stark. The large pretrained models can produce coherent text, code, images, speech and video at a scale and accuracy not possible before.

- **Enable low-friction experimentation.** The past year has seen an influx of foundation models, along with smaller, pretrained domain-specific models built from them. Most of these are available as cloud APIs or open-source projects, further reducing the time and cost to experiment.

- **Have accelerated AI innovation with massive model sizes.** Examples include OpenAI's GPT-4; Google's AI's PaLM; Google DeepMind's Gopher and Chinchilla; Meta AI's LLaMA; and Alibaba's M6. In addition, companies such as Hugging Face, Stability AI and EleutherAI have open-sourced their models.

Obstacles

Foundation models:

- **Do not deliver perfect results.** Although a significant advance, foundation models still require careful training and guardrails. Because of their training methods and black-box nature, they can deliver unacceptable results or hallucinations. They also can propagate downstream any bias or copyright issues in the datasets.

- **Require appropriate skills and talent.** As with all AI solutions, the end result depends on the skills, knowledge and talent of the trainers, particularly for prompt engineering and fine-tuning.

- **Expand to impractical sizes.** Large models are up to billions or trillions of parameters. They are impractically large to train for most organizations because of the necessary compute resources, which can make them expensive and ecologically unfriendly.

- **Concentrate power.** These models have been mostly built by the largest technology companies with huge R&D investments and significant AI talent, resulting in a concentration of power among a few large, deep-pocketed entities. This situation may create a significant imbalance in the future.

User Recommendations

- **Create a strategy document** that outlines the benefits, risks, opportunities and execution plans for these models in a collaborative effort.

- **Plan to introduce foundation models into existing speech, text or coding programs.** If you have any older language processing systems, moving to a transformer-based model could significantly improve performance. One example might be a text interpretation, where transformers can interpret multiple ideas in a single utterance. This shift in approach can significantly advance language interfaces by reducing the number of interactions.

- **Start with models that have superior ecosystem support,** have adequate enterprise guardrails around security and privacy, and are more widely deployed.

- **Explore new use cases,** such as natural language inference, sentiment analysis or natural-language-based enterprise search, where the models can significantly improve both accuracy and time to market.

- **Designate an incubation team** to monitor industry developments, communicate the art of the possible, experiment with BUs and share valuable lessons learned companywide.

Sample Vendors

Alibaba Group; Amazon; Baidu; Cohere; Google; Hugging Face; IBM; Microsoft; OpenAI; Stability AI

Generative AI

Analysis By: Svetlana Sicular, Brian Burke

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Generative AI technologies can generate new derived versions of content, strategies, designs and methods by learning from large repositories of original source content. Generative AI has profound business impacts, including on content discovery, creation, authenticity and regulations; automation of human work; and customer and employee experiences.

Why This Is Important

Generative AI exploration is accelerating, thanks to the popularity of Stable Diffusion, Midjourney, ChatGPT and large language models. End-user organizations in most industries aggressively experiment with generative AI. Technology vendors form generative AI groups to prioritize delivery of generative-AI-enabled applications and tools. Numerous startups have emerged in 2023 to innovate with generative AI, and we expect this to grow. Some governments are evaluating the impacts of generative AI and preparing to introduce regulations.

Business Impact

Most technology products and services will incorporate generative AI capabilities in the next 12 months, introducing conversational ways of creating and communicating with technologies, leading to their democratization. Generative AI will progress rapidly in industry verticals, scientific discovery and technology commercialization. Sadly, it will also become a security and societal threat when used for nefarious purposes. Responsible AI, trust and security will be necessary for safe exploitation of generative AI.

Drivers

- The hype around generative AI is accelerating. Currently, ChatGPT is the most hyped technology. It relies on generative foundation models, also called “transformers.”

- New foundation models and their new versions, sizes and capabilities are rapidly coming to market. Transformers keep making an impact on language, images, molecular design and computer code generation. They can combine concepts, attributes and styles, creating original images, video and art from a text description or translating audio to different voices and languages.

- Generative adversarial networks, variational autoencoders, autoregressive models and zero-/one-/few-shot learning have been rapidly improving generative modeling while reducing the need for training data.

- Machine learning (ML) and natural language processing platforms are adding generative AI capabilities for reusability of generative models, making them accessible to AI teams.

- Industry applications of generative AI are growing. In healthcare, generative AI creates medical images that depict disease development. In consumer goods, it generates catalogs. In e-commerce, it helps customers “try on” makeup and outfits. In manufacturing, quality inspection uses synthetic data. In semiconductors, generative AI accelerates chip design. Life sciences companies apply generative AI to speed up drug development. Generative AI helps innovate product development through digital twins. It helps create new materials targeting specific properties to optimize catalysts, agrochemicals, fragrances and flavors.

- Generative AI reaches creative work in marketing, design, music, architecture and content. Content creation and improvement in text, images, video and sound enable personalized copywriting, noise cancellation and visual effects in videoconferencing.

- Synthetic data draws enterprises’ attention by helping to augment scarce data, mitigate bias or preserve data privacy. It boosts the accuracy of brain tumor surgery.

- Generative AI will disrupt software coding. Combined with development automation techniques, it can automate up to 30% of the programmers’ work.

Obstacles

- Democratization of generative AI uncovers new ethical and societal concerns. Government regulations may hinder generative AI research. Governments are currently soliciting input on AI safety measures.
- Hallucinations, factual errors, bias, a black-box nature and inexperience with a full AI life cycle preclude the use of generative AI for critical use cases.
- Reproducing generative AI results and finding references for information produced by general-purpose LLMs will be challenging in the near term.
- Low awareness of generative AI among security professionals causes incidents that could undermine generative AI adoption.
- Some vendors will use generative AI terminology to sell subpar “generative AI” solutions.
- Generative AI can be used for many nefarious purposes. Full and accurate detection of generated content, such as deepfakes, will remain challenging or impossible.
- The compute resources for training large, general-purpose foundation models are heavy and not affordable to most enterprises.
- Sustainability concerns about high energy consumption for training generative models are rising.

User Recommendations

- Identify initial use cases where you can improve your solutions with generative AI by relying on purchased capabilities or partnering with specialists. Consult vendor roadmaps to avoid developing similar solutions in-house.
- Pilot ML-powered coding assistants, with an eye toward fast rollouts, to maximize developer productivity.
- Use synthetic data to accelerate the development cycle and lessen regulatory concerns.
- Quantify the advantages and limitations of generative AI. Supply generative AI guidelines, as it requires skills, funds and caution. Weigh technical capabilities with ethical factors. Beware of subpar offerings that exploit the current hype.
- Mitigate generative AI risks by working with legal, security and fraud experts. Technical, institutional and political interventions will be necessary to fight AI's adversarial impacts. Start with data security guidelines.
- Optimize the cost and efficiency of AI solutions by employing composite AI approaches to combine generative AI with other AI techniques.

Sample Vendors

Adobe; Amazon; Anthropic; Google; Grammarly; Hugging Face; Huma.AI; Microsoft; OpenAI; Schrödinger

Gartner Recommended Reading

Innovation Insight for Generative AI

Emerging Tech Roundup: ChatGPT Hype Fuels Urgency for Advancing Conversational AI and Generative AI

Emerging Tech: Venture Capital Growth Insights for Generative AI

Emerging Tech: Generative AI Needs Focus on Accuracy and Veracity to Ensure Widespread B2B Adoption

ChatGPT Research Highlights

Synthetic Data

Analysis By: Arun Chandrasekaran, Anthony Mullen, Alys Woodward

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Synthetic data is a class of data that is artificially generated rather than obtained from direct observations of the real world. Synthetic data is used as a proxy for real data in a wide variety of use cases including data anonymization, AI and machine learning development, data sharing and data monetization.

Why This Is Important

A major problem with AI development today is the burden involved in obtaining real-world data and labeling it. This time-consuming and expensive task can be remedied with synthetic data. Additionally, for specific use-cases like training models for autonomous vehicles, collecting real data for 100% coverage of edge cases is practically impossible. Furthermore, synthetic data can be generated without personally identifiable information (PII) or protected health information (PHI), making it a valuable technology for privacy preservation.

Business Impact

Adoption is increasing across various industries. Gartner predicts a massive increase in adoption as synthetic data:

- Avoids using PII when training machine learning (ML) models via synthetic variations of original data or synthetic replacement of parts of data.
- Reduces cost and saves time in ML development.
- Improves ML performance as more training data leads to better outcomes.
- Enables organizations to pursue new use cases for which very little real data is available.
- Is capable of addressing fairness issues more efficiently.

Drivers

- In healthcare and finance, buyer interest is growing as synthetic tabular data can be used to preserve privacy in AI training data.
- To meet increasing demand for synthetic data for natural language automation training, especially for chatbots and speech applications, new and existing vendors are bringing offerings to market. This is expanding the vendor landscape and driving synthetic data adoption.
- Synthetic data applications have expanded beyond automotive and computer vision use cases to include data monetization, external analytics support, platform evaluation and the development of test data.
- Increasing adoption of AI simulation techniques is accelerating synthetic data.
- There is an expansion to other data types. While tabular, image, video, text and speech applications are common, R&D labs are expanding the concept of synthetic data to graphs. Synthetically generated graphs will resemble, but not overlap the original. As organizations begin to use graph technology more, we expect this method to mature and drive adoption.
- The explosion of innovation in AI foundation models is boosting synthetic data creation. These models are becoming more accessible and more accurate.

Obstacles

- Synthetic data can have bias problems, miss natural anomalies, be complicated to develop, or not contribute any new information to existing, real-world data.
- Data quality is tied to the model that develops the data.
- Synthetic data generation methodologies lack standardization.
- Completeness and realism are highly subjective with synthetic data.
- Buyers are still confused over when and how to use the technology due to lack of skills.
- Synthetic data can still reveal a lot of sensitive details about an organization, so security is a concern. An ML model could be reverse-engineered via active learning. With active learning, a learning algorithm can interactively query a user (or other information sources) to label new data points with the desired outputs, meaning learning algorithms can actively query the user or teacher for labels.
- If fringe or edge cases are not part of the seed dataset, they will not be synthetized. This means the handling of such borderline cases must be carefully accommodated.
- There may be a level of user skepticism as data may be perceived to be “inferior” or “fake.”

User Recommendations

- Identify areas in your organization where data is missing, incomplete or expensive to obtain, and is thus currently blocking AI initiatives. In regulated industries, such as healthcare or finance, exercise caution and adhere to rules.
- Use synthetic variations of the original data, or synthetic replacement of parts of data, when personal data is required but data privacy is a requirement.
- Educate internal stakeholders through training programs on the benefits and limitations of synthetic data and institute guardrails to mitigate challenges such as user skepticism and inadequate data validation.
- Measure and communicate the business value, success and failure stories of synthetic data initiatives.

Sample Vendors

Anonos (Statice); Datagen; Diveplane; Gretel; Hazy; MOSTLY AI; Neuromation; Rendered.ai; Tonic.ai; YData

Gartner Recommended Reading

Innovation Insight for Synthetic Data

Innovation Insight for Generative AI

Data Science and Machine Learning Trends You Can't Ignore

Cool Vendors in Data-Centric AI

Case Study: Enable Business-Led Innovation with Synthetic Data (Fidelity International)

Edge AI

Analysis By: Eric Goodness

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Edge AI refers to the use of AI techniques embedded in non-IT products, IoT endpoints, gateways and edge servers. It spans use cases for consumer, commercial and industrial applications, such as autonomous vehicles, enhanced capabilities of medical diagnostics and streaming video analytics. While predominantly focused on AI inference, more sophisticated systems may include a local training capability to provide optimization of the AI models at the edge.

Why This Is Important

Many edge computing use cases are latency-sensitive and data-intensive, and require an increasing amount of autonomy for local decision making. This creates a need for AI-based applications in a wide range of edge computing and endpoint solutions. Examples include real-time analysis of edge data for predictive maintenance and industrial control, inferences and decision support where connectivity is unreliable, or video analytics for real-time interpretation of video.

Business Impact

The business benefits of deploying edge AI include:

- Real-time data analysis and decision intelligence
- Improved operational efficiency, such as manufacturing visual inspection systems that identify defects, wasted motion, waiting, and over- or underproduction
- Enhanced customer experience, through feedback from AI embedded within products
- Connectivity cost reduction, with less data traffic between the edge and the cloud
- Persistent functions and solution availability, irrespective of network connectivity
- Reduced storage demand, as only prioritized data is passed on to core systems
- Preserved data privacy at the endpoint

Drivers

Overall, edge AI has benefited from improvements in the capabilities of AI. This includes:

- The maturation of machine learning operationalization (MLOps) and ModelOps tools and processes support ease of use across a broader set of features that span the broader MLOps functions. Initially, many companies came to market with a narrowcast focus on model compression.
- The improved performance of combined ML techniques and an associated increase in data availability (such as time-series data from industrial assets).

Business demand for new and improved outcomes solely achievable from the use of AI at the edge, which include:

- Reducing full-time equivalents with vision-based solutions used for surveillance or inspections.
- Improving manufacturing production quality by automating various processes.
- Optimizing operational processes across industries.
- New approaches to customer experience, such as personalization on mobile devices or changes in retail from edge-based smart check-out points of sale.

Additional drivers include:

- **Increasing number of users upgrading legacy systems and infrastructure in "brownfield" environments.** By using MLOps platforms, AI software can be hosted within an edge computer or a gateway (aggregation point) or embedded within a product with the requisite compute resources. An example of this is AI software deployed (TinyML) deployed to automotive or agricultural equipment to enhance asset monitoring and maintenance.
- **More manufacturers embedding AI in the endpoint as an element of product servitization.** In this architecture, the IoT endpoints, such as in automobiles, home appliances or commercial building infrastructure, are capable of running AI models to interpret data captured by the endpoint and drive some of the endpoints' functions. In this case, the AI is trained and updated on a central system and deployed to the IoT endpoint. Examples of the use of embedded (edge) AI are medical wearables, automated guided vehicles and other robotic products that possess some levels of intelligence and autonomy.
- **Rising demand for R&D in training decentralized AI models at the edge for adaptive AI.** These emerging solutions are driven by explicit needs such as privacy preservation or the requirement for machines and processes to run in disconnected (from the cloud) scenarios. Such models enable faster response to changes in the environment, and provide benefits in use cases such as responding to a rapidly evolving threat landscape in security operations.

Obstacles

- Edge AI is constrained by the application and design limitations of the equipment deployed; this includes form factor, power budget, data volume, decision latency, location and security requirements.
- Systems deploying AI techniques can be nondeterministic. This will impact applicability in certain use cases, especially where safety and security requirements are important.
- The autonomy of edge AI-enabled solutions, built on some ML and deep learning techniques, often presents questions of trust, especially where the inferences are not readily interpretable or explainable. As adaptive AI solutions increase, these issues will increase if initially identical models deployed to equivalent endpoints subsequently begin to evolve diverging behaviors.
- The lack of quality and sufficient data for training is a universal challenge across AI usage.
- Deep learning in neural networks is a compute-intensive task, often requiring the use of high-performance chips with corresponding high-power budgets. This can limit deployment locations, especially where small form factors and lower-power requirements are paramount.

User Recommendations

- Determine whether the use of edge AI provides adequate cost-benefit improvements, or whether traditional centralized data analytics and AI methodologies are adequate and scalable.
- Evaluate when to consider AI at the edge versus a centralized solution. Good candidates for edge AI are applications that have high communications costs, are sensitive to latency, require real-time responses or ingest high volumes of data at the edge.
- Assess the different technologies available to support edge AI and the viability of the vendors offering them. Many potential vendors are startups that may have interesting products but limited support capabilities.
- Use edge gateways and servers as the aggregation and filtering points to perform most of the edge AI and analytics functions. Make an exception for compute-intensive endpoints, where AI-based analytics can be performed on the devices themselves.

Sample Vendors

Akira AI; Edge Impulse; Falkonry; Imagimob; Litmus; MicroAI; Modzy; Octonion Group; Palantir

Gartner Recommended Reading

Building a Digital Future: Emergent AI Trends

Emerging Technologies: Neuromorphic Computing Impacts Artificial Intelligence Solutions

Emerging Technologies: Edge Technologies Offer Strong Area of Opportunity — Adopter Survey Findings

Emerging Tech Impact Radar: Edge AI

ModelOps

Analysis By: Joe Antelmi, Erick Brethenoux, Soyeb Barot

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Model operationalization (ModelOps) is primarily focused on the end-to-end governance and life cycle management of advanced analytics, AI and decision models (models based on machine learning [ML], knowledge graphs, rules, optimization, linguistics, agents and others).

Why This Is Important

ModelOps helps companies challenged in standardizing, scaling and augmenting their analytics and AI initiatives that leverage a combination of statistical and ML models. It helps organizations to move their models from the lab environments into production. MLOps primarily focuses on monitoring and governance of ML models, while ModelOps assists in the operationalization and governance of all advanced analytics, decision and AI models (including ML models).

Business Impact

ModelOps delivers business impact in multiple ways. As a practice, it:

- Lays down the foundation for the management of various knowledge representation models, reasoning capabilities and composite model integration.
- Augments the ability to manage decision models and integrate multiple analytics techniques for robust decision making.
- Ensures collaboration among a wider business, development and deployment community, and the ability to correlate analytics and model outcomes with business KPIs.

Drivers

- As the number of advanced analytics, AI and decision models at organizations increase, and as projects become more complicated, organizations will have to manage different types of prepackaged or custom-made models. All will require different operationalization and governance procedures, especially when they are built from scratch.
- Organizations want to be more agile and responsive to changes within their advanced analytics and AI pipelines, not just with models, but also with data, application and infrastructure.
- The operationalization aspects of ML models are not new, but they are in their early stages. However, with ModelOps, the functionalities provided by MLOps are now extended to other non-ML models.
- ModelOps provides a framework to separate responsibilities across various teams for how models (including generative AI, foundational models, analytics, ML, physical, simulation, symbolic, etc.) are built, tested, deployed and monitored across different environments (for example, development, test and production). This enables better productivity and collaboration, and it lowers failure rates.
- There's a need to create resilient and adaptive systems that use a combination of various analytical techniques for decision support, augmentation and automation.
- There is a wide range of risk management concerns across different models — drift, bias, explainability and integrity — that ModelOps helps address.

Obstacles

- Organizations using different types of models in production often don't realize that for some analytics, decision and AI models (rule-based, agent-based, graph, generative AI or simulation models), end-to-end governance and management capabilities can and need to be expanded further.
- Not all analytical techniques currently benefit from mature operationalization methods. Because the spotlight has been on ML techniques, MLOps benefits from a more evolved AI practice, but some models, like agent-based modeling, rule-based models and optimization techniques, require more attention in ModelOps practices and platforms. The creation of applications that leverage generative AI has increased the focus of integrating ModelOps with generative and foundational models, also known as LLMOps in the industry.
- The lack of knowledge relevant to leveraging multiple analytics and AI techniques could prevent organizations from considering the techniques particularly suited to solving specific problems.
- Organizations that are siloed create redundancy in effort with respect to operationalization.

User Recommendations

- Leverage different analytics and AI techniques to increase the success rate of data and analytics initiatives.
- Utilize ModelOps best practices across data, models and applications to ensure transition, reduce friction and increase value generation.
- Extend the skills of ML experts to operationalize a wider range of models. Recruit/upskill additional AI experts to also cover graph analytics, optimization or other required techniques for composite AI
- Establish a culture that encourages collaboration between development and deployment teams, and empowers teams to make decisions to automate, scale and bring stability to the analytics and AI pipeline.
- Collaborate with data management and software engineering teams to scale ModelOps. Offloading operationalization responsibilities to multiple teams enables increased ModelOps specialization and sophistication across the ecosystem of complex AI-enabled applications.
- Optimize the adaptability and efficiency of your AI projects by considering a composite AI approach — integrating various AI techniques to solve business problems.

Sample Vendors

DataRobot; Datatron; IBM; McKinsey & Company (Iguazio); ModelOp; Modzy; SAS; Subex; Valohai; Verta

Gartner Recommended Reading

Top 5 Priorities for Managing AI Risk Within Gartner's MOST Framework

Market Guide for AI Trust, Risk and Security Management

Use Gartner's MLOps Framework to Operationalize Machine Learning Projects

A Mandate for MLOps, ModelOps and DevOps Coordination

Toolkit: Delivery Metrics for DataOps, Self-Service Analytics, ModelOps and MLOps

Sliding into the Trough

Knowledge Graphs

Analysis By: Afraz Jaffri

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Knowledge graphs are machine-readable representations of the physical and digital worlds. They include entities (people, companies, digital assets) and their relationships, which adhere to a graph data model — a network of nodes (vertices) and links (edges/arcs).

Why This Is Important

Knowledge graphs capture information about the world in an intuitive way yet are still able to represent complex relationships. Knowledge graphs act as the backbone of a number of products, including search, smart assistants and recommendation engines. Knowledge graphs support collaboration and sharing, exploration and discovery, and the extraction of insights through analysis. Generative AI models can be combined with knowledge graphs to add trusted and verified facts to their outputs.

Business Impact

Knowledge graphs can drive business impact in a variety of different settings, including:

- Digital workplace (e.g., collaboration, sharing and search)
- Automation (e.g., ingestion of data from content to robotic process automation)
- Machine learning (e.g., augmenting training data)
- Investigative analysis (e.g., law enforcement, cybersecurity or financial transactions)
- Digital commerce (e.g., product information management and recommendations)
- Data management (e.g., metadata management, data cataloging and data fabric)

Drivers

- The need to complement AI/ML methods that detect only patterns in data (such as the current generation of foundation models) with the explicit knowledge, rules and semantics provided by knowledge graphs.
- Increasing awareness of the use of knowledge graphs in consumer products and services, such as smart devices and voice assistants, chatbots, search engines, recommendation engines, and route planning.
- The emerging landscape of Web3 applications and the need for data access across trust networks, leading to the creation of decentralized knowledge graphs to build immutable and queryable data structures.
- Improvements in graph DBMS technology that can handle the storage and manipulation of graph data structures at scale. These include PaaS offerings that take away the complexity of provisioning and optimizing hardware and infrastructure.
- The desire to make better use of unstructured data held in documents, correspondence, images and videos, using standardized metadata that can be related and managed.
- The need to manage the increasing number of data silos where data is often duplicated, and where meaning, usage and consumption patterns are not well-defined.
- The use of graph algorithms and machine learning to identify influencers, customer segments, fraudulent activity and critical bottlenecks in complex networks.

Obstacles

- Awareness of knowledge graph use cases is increasing, but business value and relevance are difficult to capture in the early implementation stages.
- Moving knowledge graph models from prototype to production requires engineering and system integration expertise. Methods to maintain knowledge graphs as they scale — to ensure reliable performance, handle duplication and preserve data quality — remain immature.
- The graph DBMS market is fragmented along three properties: type of data model (RDF or property), implementation architecture (native or multimodal) and optimal workload (operational or analytical). This fragmentation continues to cause confusion and hesitation among adopters.

- Organizations want to enable the ingestion, validation and sharing of ontologies and data relating to entities (such as geography, people, events). However, making internal data interoperable with external knowledge graphs is a challenge.
- In-house expertise, especially among SMEs, is lacking, and identifying third-party providers is difficult. Often, expertise resides with vendors of graph technologies.

User Recommendations

- **Create a working group of knowledge graph practitioners and sponsors** by assessing the skills of D&A leaders and practitioners and business domain experts. Highlight the obstacles to dependable and efficient data delivery for analytics and AI, and articulate how knowledge graphs can remove them.
- **Run a pilot to identify use cases that need custom-made knowledge graphs.** The pilot should deliver not only tangible value for the business, but also learning and development for D&A staff.
- **Create a minimum viable subset that can capture the information of a business domain to decrease time to value.** Assess the data, both structured and unstructured, needed to feed a knowledge graph, and follow Agile development principles.
- **Utilize vendor and service provider expertise** to validate use cases, educate stakeholders and provide an initial knowledge graph implementation.
- **Include knowledge graphs within the scope of D&A governance and management.** To avoid perpetuating data silos, investigate and establish ways for multiple knowledge graphs to interoperate and extend toward a data fabric.

Sample Vendors

Cambridge Semantics; Diffbot; eccenca; Neo4j; Ontotext; Stardog; TigerGraph; TopQuadrant; Trace Labs

Gartner Recommended Reading

How to Build Knowledge Graphs That Enable AI-Driven Enterprise Applications

3 Ways to Enhance AI With Graph Analytics and Machine Learning

Working With Graph Data Stores

AI Maker and Teaching Kits

Analysis By: Eric Hunter, Annette Jump

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Artificial intelligence (AI) maker and teaching kits are applications and software development kits (SDKs) that abstract data science platforms, frameworks, analytic libraries and devices to enable software engineers to incorporate AI into new or existing applications. Maker kits also emphasize teaching new skills and best practices for integration between software and devices for engineers — some of which also include hardware devices.

Why This Is Important

AI maker kits package developer-friendly APIs and SDKs while often complemented by custom hardware devices (such as cameras, musical instruments, speakers or vehicles). These offerings encourage platform developer adoption while educating developers around new AI capabilities and libraries.

Business Impact

The demand for AI is significant and is increasing at a rate beyond what experienced data scientists can meet alone — with many software engineering teams leading AI development and use cases. Also, the number of sensors and data-centric enablers for AI use cases is rapidly growing. These offerings will equip software developers to become a key contingent for AI development and implementation. AI maker kits will also continue to reduce adoption barriers in the deployment of AI capabilities for software engineers and citizen data scientists.

Drivers

- As the demand for more proficient data scientists rises, the adoption of AI maker and teaching kits will continue to increase.
- Vendors of AI hardware need an educated customer base, which can be fostered through the provision of AI maker kits.
- While largely focusing on developer upskilling or proofs of concept (POCs), these kits have also driven new vendor innovations, targeting mainstream enterprise use cases. Illustrative examples include Google Coral, Microsoft Azure OpenAI Service and NVIDIA Omniverse Avatar Cloud Engine (ACE).
- Within many kits, developers can deploy prebuilt models and, optionally, update those models from cloud services at model runtime.
- AI maker and teaching kits accelerate the incorporation and deployment of AI services into your applications.
- Convergence in deployment for language, translation, vision, machine learning (ML) and cognitive search kits will support new AI use cases in the digital enterprise. Google's Multitask Unified Model (MUM) initiative is a good example of such convergence.
- Continued demand for organizational outcomes focusing on computer vision, natural language and other AI-aligned capabilities fit well with AI maker kit offerings.
- Specialist hardware linked to AI systems is more easily demonstrated, and exploited, through AI maker kits, such as the desktop-scale robot arm from Rotrics.
- Maker kit paths to production-caliber offerings will continue to mature as new capabilities are integrated into market offerings.

Obstacles

- Vendor kit and hardware innovations have stagnated, most traction is around software API and SDK innovations.
- Vendor offerings require distinct deployment considerations and have varied feature coverage differences, but we expect greater consistency in the future.
- Data scale and management strategies can be overlooked as ideas move beyond a POC stage.
- Kits support only a limited set of native use cases, such as computer vision, image recognition, labeling, natural language and text analytics.
- Market offerings are typically mutually exclusive in terms of the use cases supported, usually being singular (i.e., a computer vision kit and a kit supporting natural language processing [NLP] have no shared components or platforms).
- Market offerings do not follow a consistent set of standards, and kits have an inconsistent level of support/capabilities for production-ready use cases. Some support scaling development concepts to full-scale production use cases, while others offer no path from development-only scenarios.

User Recommendations

- Leverage maker kits to upskill developer knowledge and skills, which can translate to present and future enterprise needs that may directly or indirectly relate to kit-specific use cases.
- Carefully evaluate and stress-test employed maker kit offerings, along with fully understanding the going concern support for each specific offering.
- Abstract adopted vendor development kit offerings where possible to minimize portability constraints and lock-in, and prioritize vendor interoperability.
- Actively participate in commercial and open-source communities to influence longer-term roadmaps.
- Ensure deployed capabilities are aligned to direct end-user benefits that cannot be easily achieved without AI.
- Adopt offerings in alignment with larger organizational development standards and strategies.

Sample Vendors

Amazon Web Services; Google; Intel; Microsoft; NVIDIA; Pantech Prolabs India; Rotrics; Samsung Electronics

Gartner Recommended Reading

Emerging Tech Impact Radar: Artificial Intelligence

Emerging Technologies: Top Use Cases for Smart Robots to Lead the Way in Human Augmentation

Applying AI — A Framework for the Enterprise

What Is Artificial Intelligence? Ignore the Hype; Here's Where to Start

Improve Computer Vision Use Cases With Standardized Implementation Patterns

Autonomous Vehicles

Analysis By: Jonathan Davenport

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Autonomous vehicles use various onboard sensing and localization technologies, such as lidar, radar, cameras, global navigation satellite system (GNSS) and map data, in combination with AI-based decision making, to drive without human supervision or intervention. Autonomous vehicle technology is being applied to passenger vehicles, buses and trucks, as well as for specific use cases such as mining and agricultural tractors.

Why This Is Important

Autonomous vehicles have the potential to change transportation economics, cutting operational costs and increasing vehicle utilization. In urban areas, inexpensive fares and high-quality service may reduce the need for private car ownership. Road safety will also increase, as AI systems will never be distracted, drive drunk or speed. Autonomous features on privately owned vehicles will enable productivity and recreational activities to be undertaken, while the vehicle handles the driving operations.

Business Impact

Autonomous vehicles have the potential to disrupt established automotive business models. Self-driving systems will stimulate demand for onboard computation, radically increasing the semiconductor content of vehicles. After the office and home, vehicles will become a living space (like airplanes) where digital content is both created and consumed. Over time, it is likely that fleet operators will retrain and redeploy commercial drivers to other, higher-value-adding roles within the company.

Drivers

- The formalization of regulations and standards for autonomous vehicles will aid implementation. Automated lane-keeping system (ALKS) technology has been approved by the United Nations Economic Commission for Europe (UNECE). This is the first binding international regulation for SAE Level 3 vehicle automation, with a maximum operational speed of 37 mph. To take advantage of the new regulatory landscape, automakers are beginning to announce Level 3 solutions. Honda is the first company to announce a commercially available ALKS equipped vehicle, though only 100 will be produced.

- Other companies are quickly following, with Mercedes-Benz being the first automotive manufacturer worldwide to secure internationally valid system approval and has launched in Germany. Its Level 3 solution has secured approval from the state of Nevada and an application to enable cars to drive autonomously in California has also been made.

- In China Changan, Great Wall Motor and Xpeng have announced Level 3 systems. Other global automakers are following suit. Hyundai's new Genesis G90 and the Kia EV9 vehicles will come equipped with a Level 3 Highway Driving Pilot (HDP) function.

- This signals that the autonomous vehicle market is most likely to evolve gradually from ADAS systems to higher levels of autonomy on passenger vehicles, rather than seeing a robotaxi-based revolution. This will require flexible vehicle operational design domains (ODDs). Progress is being made by companies like Mobileye who's perception system was developed on the roads of Israel, but required minimal retraining to perform well in diverse cities like Munich and Detroit.

- The most compelling business case for autonomous vehicles relates to self-driving trucks. Driver pay is one of the largest operating costs for fleets associated with a commercial truck, plus goods can be transported much faster to their destination because breaks are no longer necessary. The Aurora Driver product is now at a "feature complete" stage, with a plan to launch a "middle-mile" driverless truck service at the end of 2024.

Obstacles

- Designing an AI system that is capable of driving a vehicle is hugely complex. As a result, the cost of bringing a commercial autonomous vehicle to market has been greater than companies could have previously envisioned, requiring significant investments to be made.
- When autonomous vehicles are commercially deployed, autonomous vehicle developers, not the human occupants, will be liable for the autonomous operations of the vehicle. This raises important issues, should a vehicle be involved in an accident, and the need for associated insurance.
- Challenges increasingly include regulatory, legal and societal considerations, such as permits for operation and the effects of human interactions.

Analyst Notes:

- Volvo's EX90 vehicles are being deployed with hardware-ready for unsupervised autonomous driving (including a lidar from Luminar), despite the self-driving software not being ready for deployment. Volvo plans to deploy an over-the-air software to move capability from Level 2 ADAS system to Level 3 in the future.
- Despite continued improvements in Level 4 autonomous vehicle perception algorithms and broader self-driving systems used for mobility use cases (such as robotaxis), driverless operations have not scaled to different cities quickly. Waymo — one of the early leaders of operations without a safety driver — has struggled to expand outside of Arizona.
- Slow progress saw Ford and VW pull their investments in Argo AI at the end of 2022, causing the joint venture's operation to close. VW had invested approximately 2 billion Euros in the company.
- Likewise Pony.ai's permit to test driverless vehicles in California was suspended after an accident and in San Francisco, Cruise's autonomous operations led to traffic disruptions — local transit officials cited 92 unique incidents between 29th May and 31st December 2022.

User Recommendations

Governments must:

- Craft national legislation to ensure that autonomous vehicles can safely coexist with a traditional vehicle fleet as well as a framework for their approval and registration.
- Work closely with autonomous vehicle developers to ensure that first responders can safely respond to road traffic and other emergencies and self-driving vehicles don't obstruct or hinder activities.

Autonomous mobility operators should:

- Support consumer confidence in autonomous vehicle technology by remaining focused on safety and an accident-free road environment.

Traditional fleet operators looking to adopt autonomous technology into their fleets should:

- Minimize the disruptive impact on driving jobs (bus, taxi and truck drivers) by developing policies and programs to train and migrate these employees to other roles.

Automotive manufacturers should:

- Instigate a plan for how higher levels of autonomy can be deployed to vehicles being designed and manufactured to future-proof vehicle purchases and enable future functions-as-a-service revenue streams.

Sample Vendors

Aurora; AutoX; Baidu; Cruise; Mobileye; NVIDIA; Oxbotica; Pony.ai; Waymo; Zoox

Gartner Recommended Reading

Emerging Tech Impact Radar: Autonomous Vehicles, 2022

Lessons From Mining: 4 Autonomous Thing Benefit Zones for Manufacturers

Forecast Analysis: Autonomous Vehicle Net Additions, Internet of Things, Worldwide

Tech Providers 2025: Product Leaders Must Strategize to Win in the Evolving Robotaxi Ecosystem

Climbing the Slope

Cloud AI Services

Analysis By: Van Baker, Bern Elliot

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

AI cloud services provide AI model building tools, APIs for prebuilt services and associated middleware that enable the building/training, deployment and consumption of machine learning (ML) models running on prebuilt infrastructure as cloud services. These services include vision and language services and automated ML to create new models and customize prebuilt models.

Why This Is Important

The use of AI cloud services continues to increase. Vendors have introduced additional services and solutions with fully integrated MLOps pipelines. The addition of low-code tools has added to ease of use. Applications regularly use AI cloud services in language, vision and automated ML to automate and accelerate business processes. Developers are aware of these offerings, and are using both prebuilt and customized ML models in applications.

Business Impact

The impact of AI extends to the applications that enable business, allowing developers and data scientists to enhance the functionality of these applications. The desire for data-driven decisions in business is driving the incorporation of forecasts and next best actions, including automation of many workflows. AI cloud services enable the embedding of advanced machine learning models in applications that are used to run the day-to-day business operations.

Drivers

- **Opportunities to capitalize on new insights.** The wealth of data from both internal and third-party sources delivers insights that enable data-driven decision intelligence.

- **Support demand for conversational interactions.** The emergence of generative AI and large language models facilitates conversationally enabled applications.
- **To meet business key performance indicators (KPIs).** There is a mandate for businesses to automate processes to improve accuracy, improve responsiveness and reduce costs by deploying both AI and ML models.
- **Reduced barriers of entry.** The ability to do zero-shot learning and model fine-tuning has reduced the need for large quantities of data to train models. Access for developers and citizen data scientists to AI and ML services due to the availability of API callable cloud-hosted services will expand the use of AI.
- **AutoML as an enabler for custom development.** Use of automated ML to customize packaged services to address specific needs of the business is increasing.
- **A wide range of AI cloud services.** AI cloud services from a range of specialized providers in the market, including orchestration layers to streamline deployment of solutions, are available.
- **Emerging AI model marketplaces.** New marketplaces should help developers adopt these techniques through AI cloud services.

Obstacles

- **Lack of understanding** by developers and citizen data scientists about how to adapt these services to specific use cases.
- **Pricing models** for AI cloud services that are usage-based presents a risk for businesses as the costs associated with use of these services can accrue rapidly.
- **Increased need** for packaged solutions that utilize multiple services for developers and citizen data scientists.
- **Lack of marketplaces** for prebuilt ML models that can be adapted for specific enterprise use cases.
- **Continuing need for ModelOps** tools that enable integration of AI and ML models into applications.
- **Lack of skills** for developers to effectively implement these services in a responsible manner.

User Recommendations

- Choose customizable AI cloud services over custom models to address a range of use cases and for quicker deployment and scalability.
- Improve chances of success of your AI strategy by experimenting with AI techniques and cloud services, using the exact same dataset and selecting one that addresses requirements. Consider using an A/B testing approach.
- Use AI cloud services to build less complex models, giving the benefit of more productive AI while freeing up your data science assets for higher-priority projects.
- Empower non-data-scientists with features such as automated algorithm selection, dataset preparation and feature engineering for project elements. Leverage existing expertise on operating cloud services to assist technical professional teams.
- Establish a center of excellence for responsible use of AI that includes all functional areas of the business. This is especially important in light of the advances of generative AI solutions.

Sample Vendors

Alibaba; Amazon Web Services; Baidu; Clarifai; Google; H2O.ai; Huawei; IBM; Microsoft; Tencent

Gartner Recommended Reading

Critical Capabilities for Cloud AI Developer Services

Magic Quadrant for Cloud AI Developer Services

Intelligent Applications

Analysis By: Alys Woodward, Justin Tung, Stephen Emmott

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Intelligent applications utilize learned adaptation to respond autonomously to people and machines. While applications can behave intelligently, intelligent applications are intelligent. Rule-based approaches, dependent upon conditional logic are giving way to math-based training to elicit an appropriate response across a broader range of circumstances, including those that are new or unique. This enables the augmentation and automation of work across a broader range of scenarios and use cases.

Why This Is Important

Artificial intelligence (AI) is the current competitive play for enterprise applications with many technology providers now enabling AI and machine learning (ML) in their products via inbuilt, added, proxied or custom capabilities. Bringing intelligence into applications enables them to work autonomously across a wider range of scenarios with elevated quality and productivity, and reduced risk. Integrated intelligence can also support decision-making processes alongside transactional processes.

Business Impact

Three core benefits to organizations buying or augmenting intelligent applications are:

- **Automation** — They increase automated and dynamic decision making, reducing the cost and unreliability of human intervention, and improve the effectiveness of business processes.
- **Augmentation** — They increase the quality of dynamic decision making based on context and risk, whether automated or via improved decision support.
- **Contextualization** — Applications can adapt to the context of the user or process.

Drivers

- **The hype wave for large language models (LLMs) and applications built upon them like OpenAI's ChatGPT have inspired business users to use conversational user interfaces (UIs) wherever possible.** Conversational UIs, while a part of AI, are more about improving the interface and less about adding intelligence to the application. This is despite the fact that ML is the most widespread and impactful enabler for intelligent applications because it allows features like recommendations, insights and personalization. Wider incorporation of chat-based interfaces will blur the line between interface and intelligence in an easily composable manner.

- **AI capabilities and features are increasingly being integrated** into ERP, CRM, digital workplace, supply chain and knowledge management software within enterprise application suites. LLMs can be added on top to replace interfaces.

- **Organizations are demanding more functionalities from applications,** whether built or bought, expecting them to enhance current processes for both transactions and decision making with recommendations, insights and additional information. This in turn allows vendors to deliver higher value and drive higher prices.

- **The trend toward composable business architectures** has highlighted the possibilities around delivering advanced and flexible capabilities to support, augment and automate decisions. It required an underlying data fabric and packaged capabilities to build this, but LLMs used as a composable interface layer could kick-start the ability to deliver on the composable architecture.

- **Vendor investment in intelligent applications in the form of adding intelligent components achieves higher ROI.** In the 2022 AI technology benchmark survey of intelligent applications vendors, the median number of customers (end-user organizations) adopting intelligent applications was 35%, and the median return on investment for vendors was 25x of their initial investment.

Obstacles

- **Lack of data —** Intelligent applications require access to data from a range of systems, meaning application vendors need to think about data management technology and processes outside their own solutions.
- **Adding AI adds complexity to operations —** Models have to be trained and maintained, and users must understand the latency of the data. Contextualizing insights requires business metadata.
- **Overuse of AI in marketing —** Vendors sometimes neglect the focus on business impact, which can generate a cynical response in business buyers, particularly when AI has not delivered value in the past.
- **Trust in system-generated insights —** It takes time for business users to see the benefit and trust such insights and some explainability is key.
- **The rapid rise of conversational AI UIs —** Since December 2022, the rise of ChatGPT and similar applications has sparked great interest and activity in chat interfaces, and adds the ability to compose a conversational layer on top of existing legacy applications.

User Recommendations

- Challenge your packaged software providers to outline in their product roadmaps and/or ecosystems how they are incorporating AI to add business value in the form of a range of AI technologies.
- Evaluate the architecture of your providers by considering that the best-in-class intelligent applications are built from the ground up to be constantly collecting data from other systems, with a solid data layer in the form of a data fabric.
- Prioritize investments in specialized and domain-specific intelligent applications delivered as point solutions, which help solve problem areas such as customer engagement and service, talent acquisition, collaboration, and engagement.
- Bring AI components into your composable enterprise to innovate faster and safer, to reduce costs by building reusability, and to lay the foundation for business-IT partnerships. Remain aware of what makes AI different, particularly how to refresh ML models to avert implementation and usage challenges.

Sample Vendors

ClayHR; Creatio; Eightfold AI; JAGGAER; Prevedere; Pricefx; Salesforce; Sievo; SugarCRM; Trust Science

Gartner Recommended Reading

Top Tech Provider Trend for 2023: Intelligent Applications

Emerging Tech: Intelligent Applications Vendors Invest to Differentiate — Benchmark Findings

Strategic Roadmap for the Composable Future of Applications

Data Labeling and Annotation

Analysis By: Svetlana Sicular, Alexander Linden, Anthony Mullen

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

Data labeling and annotation (DL&A) is a process where data assets are further classified, segmented, annotated and augmented to enrich data for better analytics and artificial intelligence (AI) projects. Associated services and platforms route and allocate these tasks to both internal staff and external third-party knowledge workers to optimally manage the required workflows and thus improve the quality of training data.

Why This Is Important

The need for better training data has increased to remove the bottleneck in developing AI solutions — especially those particular to generative AI and industry use cases. Given the typical lack of internal skills and systems, DL&A services and tools are often the best option (by cost, quality and availability) to provide necessary data for best AI results. Today, at least, some AI solutions would not be possible at their current levels without human-based labeling and its further automation.

Business Impact

Major impacts of DL&A are:

- Enables AI solutions where they may not have been possible before due to lack of training data
- Improves accuracy of foundation models, mostly via reinforcement learning with human feedback (RLHF)
- Adjusts generative AI outcomes to organization-specific needs
- Drives higher performance of AI solutions because of richer, annotated datasets
- Speeds times around model development — can flex to accommodate varying workloads
- Can support a reboot of metadata management practices

Drivers

Growth of investments in AI:

- **Increased exploitation of pretrained, off-the-shelf models:** When models are pretrained, data is the main means for customization and fine-tuning of the models.
- **Increased diversity of use cases:** These services can accelerate and unlock a wealth of use cases across all industries, and core competencies in natural language automation and computer vision. Vendors in the marketplace today have dedicated offerings for commerce, robotics and autonomous vehicles, retail, GIS/maps, AR/VR, agriculture, finance, manufacturing and transportation, and communications.
- **A shift from tactical to strategic offerings:** Some language-centric platforms such as insight engines, semantic AI platforms and intelligent document processing solutions connect labeling tasks to master metadata management, yielding reusable strategic assets like taxonomies, ontologies and knowledge graphs.
- **Major AI vendors offer DL&A crowdsourced and automated services** to streamline the AI process on their AI and ML platforms.

Rise of generative AI:

- Growth of large language model (LLM) offerings: Natural language technology workload outsourcing for speech, conversational AI and document labeling is a major area of growth in this market. The DL&A market is expanding beyond row data to look at corpora quality to make LLMs more performant.

- RLHF, a type of LLM fine-tuning, involves a reward model that rates how good a model's response is from the perspective of a human labeler — this guides the model's learning process. The process involves human-labeled prompts and labeling preferred responses.
- Generative AI methods allow lowering the cost of DL&A through automation. LLMs are increasingly used to extract labels from text data through zero-shot learning.

Obstacles

- **Challenger methods:** Few-shot learning, transfer learning, synthetic data, semantic platforms and data marketplaces compete for use.
- **Third-party worker quality:** Challenges remain around third-party knowledge workers' quality and security to annotate the data, somewhat ameliorated by the development of reputation systems and prequalification tests.
- **No consolidation of AI-task-outsourcing marketplaces:** The translation ecosystem, the gig economy, and data labeling and annotation are as yet not a simplified, coherent "language operation" for organizations.
- **Supply outstrips demand and price points are often uneconomical for large-scale data:** Many vendors have entered this space in the last few years, and demand from buyers does not yet match supply. Pricing and business models vary considerably among providers, and buyers find it difficult to estimate costs.
- **Security concerns:** Especially for those DL&A services that bring in public crowds, many clients feel uneasy distributing certain data to virtually unknown parties.

User Recommendations

- Design development and production workflows to leverage a mixture of internal and external knowledge workers to support data labeling.
- Prioritize DL&A automation options when possible.
- Ensure the provider you choose has methods to test its pool of knowledge workers for domain expertise and measures of accuracy and quality.
- Model costs to avoid surprises by exploring and estimating the spend across the variety of business models, which range from label volumes and project-based to per annotator/seat costs.
- Allow data scientists to focus on more valuable tasks and lighten their load in classifying and annotating data by using DL&A services.
- Use vendors with real-time human-in-the-loop solutions for production systems like chatbots and recommenders to handle low-confidence thresholds, spikes in demand or access to real-time knowledge not present in the enterprise.
- Mature ad hoc tactical labeling activities to a systemwide in-the-line-of-work approach to curate ongoing master metadata schemas.

Sample Vendors

CrowdWorks; Defined.ai; Diffgram; Heartex; Isahit; Labelbox; Mindy Support; Scale AI; Snorkel AI

Gartner Recommended Reading

Best Practices for the Responsible Use of Natural Language Technologies

Market Guide for AI-Enabled Translation Services

Emerging Tech: Tech Innovators in Synthetic Data for Image and Video Data — Domain-Focused

Computer Vision

Analysis By: Nick Ingelbrecht, Shubhangi Vashisth

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Computer vision is a set of technologies that involve capturing, processing and analyzing real-world images and videos to extract meaningful, contextual information from the physical world.

Why This Is Important

Computer vision comprises a transformational collection of technologies that are essential to sensing and understanding the physical environment. Computer vision technology is driving innovation across many industries and use cases and is creating unprecedented business applications and opportunities.

Business Impact

Computer vision technologies are used across all industries and address a broad and growing range of business applications. These include physical security, retail and commercial property, automotive, robotics, healthcare, manufacturing, supply chain/logistics, banking and finance, agriculture, government, media and entertainment, and Internet of Things (IoT). Computer vision exploits the visible and nonvisible spectrum, including infrared, hyperspectral imaging, lidar, radar and ultraviolet.

Drivers

Computer vision adoption is being driven by improvements in the application of machine learning methods, tools and services, hardware processing efficiencies, and data generation and augmentation techniques:

- **New neural network architectures, models and algorithm enhancements** are steadily improving the price/performance of computer vision applications; combinations of CNNs and vision transformers are delivering leading levels of performance; model compression and chip advancement enable larger workloads to be run on edge devices.

- **The economics of computer vision are being enhanced by the growth of the market for computer vision tools and services.** These include annotation and data preparation services and automated machine learning (autoML) capabilities, reaching across computer vision data pipelines, from model development and training through to deployment and model management, maintenance, and governance.

- **The proliferation of cameras and other sensors is generating exponential increases in image data,** creating a critical and growing demand for methods to automate analysis and manage and extract value from that data. Dynamic vision systems and lower cost lidar products are opening new areas for innovation.

- **Edge-enabled cloud frameworks,** developer ecosystems, products and support are further expanding the opportunity and enabling non-experts to train and deploy their own computer vision models.

- **New business models and applications** are emerging, ranging from smartphone cameras and fun filters, through to global video content production and distribution, life-saving medical image diagnostics, autonomous vehicles, video surveillance for security, robotics and manufacturing automation.

- **Sensor fusion,** multimodal analysis, generative AI, multispectral and hyperspectral imaging are expanding the opportunities.

- **Improved reliability,** price, performance and functionality are generating compelling business value and driving adoption.

Obstacles

Enterprises struggle with how best to exploit their visual information assets and automate the analysis of exponential volumes of image data:

- High-end systems are expensive to maintain and support, and building business cases with adequate ROI is challenging.

- The computer vision market lacks independent standardization and performance benchmarks, and advanced solutions are far from being commoditized.

- Integration with existing systems is problematic due to a lack of open interfaces, off-the-shelf solutions and plug-and-play capabilities.

- Enterprises struggle to activate computer vision models in business processes and face data security and privacy risks.

- Scaling solutions is challenging due to the high levels of customization and service support needed.
- Adequate training and testing data may be hard or expensive to acquire, especially in areas where available open-source computer vision datasets are declining.
- Proprietary algorithms and patent pools deter innovation.

User Recommendations

- Assess change management impacts of computer vision projects on the organization and its people.
- Focus initially on a few small projects, using fail-fast approaches and scale the most promising systems into production using cross-disciplinary teams.
- Test production systems early in the real-world environment because lighting, color, object disposition and movement can break computer vision solutions that worked well in the development cycle.
- Build internal computer vision competencies and processes for exploiting image and video assets.
- Exploit third-party computer vision tooling and services to accelerate data preparation and reduce costs.
- Evaluate legal, regulatory, commercial and reputational risks associated with computer vision projects at the outset.
- Reduce the barrier to computer vision adoption by addressing two of the main challenges, lack of training data and costly and constrained hardware, by investing in synthetic and augmented data solutions and model compression to improve model performance and expand the range of more valuable use cases.

Sample Vendors

Amazon Web Services; Baidu; Clarifai; Deepomatic; Google; Matroid; Microsoft Azure; Tencent

Gartner Recommended Reading

Emerging Technologies: Emergence Cycle for Computer Vision

Emerging Tech: Revenue Opportunity Projection of Computer Vision

Emerging Technologies: Computer Vision Is Advancing to Be Smarter, More Actionable and on the Edge

Emerging Technologies Tool: Video Analytics Functionality Matrix

Emerging Technologies: Tech Innovators for Computer Vision

Appendixes

See the previous Hype Cycle: Hype Cycle for Artificial Intelligence, 2022

Hype Cycle Phases, Benefit Ratings and Maturity Levels

Table 2: Hype Cycle Phases

(Enlarged table in Appendix)

Phase ↓	Definition ↓
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant media and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the innovation is pushed to its limits. The only enterprises making money are conference organizers and content publishers.
Trough of Disillusionment	Because the innovation does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the innovation's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the innovation are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of Productivity.

Source: Gartner (July 2023)

Table 3: Benefit Ratings

Benefit Rating	*Definition*
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (July 2023)

Table 4: Maturity Levels

(Enlarged table in Appendix)

Maturity Levels	*Status*	*Products/Vendors*
Embryonic	In labs	None
Emerging	Commercialization by vendors Pilots and deployments by industry leaders	First generation High price Much customization
Adolescent	Maturing technology capabilities and process understanding Uptake beyond early adopters	Second generation Less customization
Early mainstream	Proven technology Vendors, technology and adoption rapidly evolving	Third generation More out-of-box methodologies
Mature mainstream	Robust technology Not much evolution in vendors or technology	Several dominant vendors
Legacy	Not appropriate for new developments Cost of migration constraints replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (July 2023)

Document Revision History

Hype Cycle for Artificial Intelligence, 2022 - 8 July 2022

Hype Cycle for Artificial Intelligence, 2021 - 29 July 2021

Hype Cycle for Artificial Intelligence, 2020 - 27 July 2020

Hype Cycle for Artificial Intelligence, 2019 - 25 July 2019

Hype Cycle for Artificial Intelligence, 2018 - 24 July 2018

Hype Cycle for Artificial Intelligence, 2017 - 24 July 2017

Hype Cycle for Smart Machines, 2016 - 21 July 2016

Hype Cycle for Smart Machines, 2015 - 24 July 2015

Hype Cycle for Smart Machines, 2014 - 18 July 2014

Recommended by the Author

Some documents may not be available as part of your current Gartner subscription.

Understanding Gartner's Hype Cycles

Tool: Create Your Own Hype Cycle With Gartner's Hype Cycle Builder

The Future of AI: Reshaping Society

Innovation Insight for Generative AI

Applying AI — A Framework for the Enterprise

AI Design Patterns for Large Language Models

Table 1: Priority Matrix for Artificial Intelligence, 2023

Benefit	*Years to Mainstream Adoption*			
↓	***Less Than 2 Years*** ↓	***2 - 5 Years*** ↓	***5 - 10 Years*** ↓	***More Than 10 Years*** ↓
Transformational	Computer Vision	Composite AI Decision Intelligence First-Principles AI Generative AI Intelligent Applications	Autonomic Systems Autonomous Vehicles Foundation Models Neuromorphic Computing Responsible AI	Artificial General Intelligence
High	Data Labeling and Annotation Edge AI	AI Maker and Teaching Kits AI TRiSM Causal AI Cloud AI Services Data-Centric AI Knowledge Graphs Prompt Engineering Synthetic Data	AI Engineering AI Simulation ModelOps Multiagent Systems Operational AI Systems Smart Robots	Neuro-Symbolic AI
Moderate				
Low				

Source: Gartner (July 2023)

Table 2: Hype Cycle Phases

Phase ↓	*Definition* ↓
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant media and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the innovation is pushed to its limits. The only enterprises making money are conference organizers and content publishers.
Trough of Disillusionment	Because the innovation does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the innovation's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the innovation are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of Productivity.

Phase	*Definition*

Source: Gartner (July 2023)

Table 3: Benefit Ratings

Benefit Rating	*Definition*
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (July 2023)

Table 4: Maturity Levels

Maturity Levels ↓	Status ↓	Products/Vendors ↓
Embryonic	In labs	None
Emerging	Commercialization by vendors Pilots and deployments by industry leaders	First generation High price Much customization
Adolescent	Maturing technology capabilities and process understanding Uptake beyond early adopters	Second generation Less customization
Early mainstream	Proven technology Vendors, technology and adoption rapidly evolving	Third generation More out-of-box methodologies
Mature mainstream	Robust technology Not much evolution in vendors or technology	Several dominant vendors
Legacy	Not appropriate for new developments Cost of migration constraints replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (July 2023)