The evolution of AI agents

Charting the evolution of AI agents

When large language models (LLMs) first emerged, most notably in the widely accessible form of ChatGPT, users were enchanted by their ability to think. They prompted these models with questions, and the models used their understanding gained from training on tons of data, and responded in a surprisingly humanlike way.

But humans can do more than just think and talk. For more than a half-century, AI researchers and philosophers have asked an obvious question: what if artificial intelligences were able to not only think and talk—what if they could also "do"? What if they were more than just isolated artificial brains, and could autonomously respond to and act upon their environment in real time?

This line of questioning resulted in the current era of agentic AI, the very beginning stages of which we are now experiencing. The history of AI is only one facet of the story of agents, with other strands including research in robotics, computation and cognitive theory, all developing in parallel, and in fits and starts across the last century.

Your subscription will be delivered in English. You will find an unsubscribe link in every newsletter. You can manage your subscriptions or unsubscribe here. Refer to our IBM Privacy Statement for more information.

Early philosophical roots (1940-1960)

Esoteric texts and science fiction literature going back hundreds of years has pondered the concepts of artificial intelligences and automata that have the ability to think and act autonomously. These early musings laid the groundwork for imagining the mechanisms that would one day be called AI agents.

The field of AI is generally considered to properly begin in the mid-20th century. Norbert Wiener's work in cybernetics, which focused on communication and control systems in both living beings and machines, introduced the idea of feedback loops.[1] These systems enabled entities to sense their environment, process information and adjust behavior accordingly. Even a very simple mechanism like

a thermostat has this ability: it senses temperature, compares it against its setting, and activates (or doesn't activate) a furnace in response. Even this can be considered an agent, albeit a simple reflex agent. Automatic thermostats have existed in rudimentary form since at least the 1600s, but now we had better language to describe them.[2]

In 1943, Warren S. McCulloch and Walter Pitts published "A Logical Calculus of the Ideas Immanent in Nervous Activity" in the *Bulletin of Mathematical Biophysics*.[3] Considered one of the seminal works in neuroscience and AI, the paper introduced the idea that the brain can be analyzed as a computational system. The now-common concept of neural networks traces back to this work.

In 1950, Alan Turing's landmark paper "Computing Machinery and Intelligence" is published in *Mind*.[4] The paper inquired into the nature of thinking machines and how their intelligence might be measured with a process now known as the "Turing Test."

Oliver Selfridge, in his 1959 paper "Pandemonium - A Paradigm for Learning," established conceptual structures that later agentic architectures would echo.[5] Pandemonium was built around "demons": small, specialized computational entities. Each demon had a narrow responsibility, and they operated in parallel. Pandemonium showed that intelligence could emerge from competition and cooperation among simple decision-making entities.

Logic and problem solving (1950-1970)

Although the philosophical groundwork for thinking machines had already been laid, it wasn't until the birth of digital computing in the 1950s that these ideas could be executed. Marvin Minsky and Dean Edmunds built one of the first artificial neural networks in 1951, the Stochastic Neural Analog Reinforcement Calculator (SNARC).[6] This was an early attempt to model learning processes in the human brain through reinforcement learning. The hardware required a network of 3000 vacuum tubes alongside synaptic weights to simulate 40 neuron-like units.

In 1955, the term "artificial intelligence" was coined in a workshop proposal titled "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence."[7] The proposal, submitted by John McCarthy of Dartmouth College, Marvin Minsky of Harvard University, Nathaniel Rochester from IBM and Claude Shannon from Bell Telephone Laboratories, was realized as a workshop which took place the following year.

Also in the following year, Allen Newell and Herbert A. Simon developed the Logic Theorist and General Problem Solver, early attempts at mimicking human problem-solving abilities.

Research continued throughout the decade, with researchers even developing rudimentary AI agents, although these agents could still not "learn" in the contemporary sense. For example, an agent could be designed to solve a puzzle by having it analyze the current state of the puzzle, comparing it with the solved state, and applying a known sequence of moves to get closer to that solved state. These agents merely executed algorithms to search through a possibility space. However, they proved that machines could exhibit behaviors that had previously been thought of as limited to human cognition.

In the 1970s, a new class of proto-agents emerged that used symbolic decision logic. These "expert systems" were created to capture the knowledge and reasoning capabilities of human experts by combining a knowledge base with an inference engine. MYCIN, developed at Stanford, was an early example of such systems.[8] MYCIN could diagnose bacterial infections and recommend antibiotics, finding success in narrow domains. However, every new piece of data had to be hand-coded by human experts, and these systems struggled when operating outside their rigid knowledge bases.

These limitations and others resulted in a period that came to be known as the "AI winter," where funding and by extension, AI research, constricted.

Paving the way for agents (1970-2010)

It would take decades for the AI winter to fully thaw, but progress was still made during this period.

Computation

In 1973, computer scientist Carl Hewitt developed the actor model, which treats an actor as the fundamental building block in concurrent computation, a form of computing where computations happen at the same time rather than sequentially.

The actor receives and sends messages, creates new actors and decides how to respond next based on its internal state. The actor model helped define agents as active entities rather than passive data structures—computation emerges from their behavior.[9] This theoretical framework is very close to how we think about AI agent functionality and architecture today.

The same can be said for object oriented programming (OOP). In the 1970s and '80s, OOP rose to prominence. In this programming paradigm, each object (fundamental building blocks of programs), has its own internal state and manages itself. Objects communicate by invoking behaviors on each other, establishing the expectation that multiple software entities coordinate through messaging rather than sharing memory.

Across the '80s and '90s, researchers began to formally use the word "agent" to describe an artificial intelligences that could "do." There were a number of theoretical agentic frameworks introduced during this period. The belief–desire–intention (BDI) model, offered a structured way to think about autonomous systems capable of navigating dynamic environments. Distributed AI explored networks of cooperating processes, while the field of mobile agents allowed software to move across machines to perform work.

Robotics

Robotics also yielded fruitful results for researchers, if not yet for practical use. As far back as 1966, robots had been technically behaving as rudimentary agents. "Shakey" was the world's first mobile robot to combine AI, sensing and logical reasoning to navigate and complete tasks in a real-world environment.

Another example is Rodney Brooks' subsumption architecture, a reactive robotic architecture emphasizing layered, decentralized behaviors that don't rely on complete world models.[10] Subsumption 'agents' perform modular behaviors within reactive perception-action loops, like today's AI agents.

These developments in robotics collectively formed the intellectual bedrock of the modern agent: a system capable of autonomy, reactivity, and proactiveness.

Despite this conceptual progress, the agent remained limited by the broader capabilities of AI. It would take decades for machine learning to develop to the point where robotic learning agents could become useful tools.

Machine learning

Computational power increased and datasets grew, and advancements in machine learning helped researchers put theory to practice.

In 1986, David Rumelhart, Geoffrey Hinton and Ronald Williams published "Learning representations by back-propagating errors," in which they described the backpropagation algorithm.[11] Backpropagation sparked a renewed interest in neural networks by providing an efficient way to train them. Short for "backward propagation of error", backpropagation is an elegant method to calculate how changes to any of the weights or biases of a neural network will affect the accuracy of model predictions.

In 1989, Yann LeCun and a team of researchers at AT&T Bell Labs applied backpropagation to image recognition using convolutional neural networks (CNNs), one of the first applications of deep learning, so-called because of the many layers of neurons used in CNNs. Deep learning models could extract hierarchical features from raw data, allowing models to perceive and interpret more complex data. Deep learning continued to develop throughout the 2000s and 2010s.

As the internet expanded, it played a critical role in AI's resurgence. All that data (images, text, videos) became the fuel for ever more sophisticated machine learning algorithms.

Agentic architectures emerge

In 1995, Russell and Norvig wrote in their massively popular *Artificial Intelligence: A Modern Approach*, that agents "operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change and create and pursue goals."[12] That same year, Woolridge and Jennings contributed to the field by defining an agent as having autonomy, social ability, reactivity and proactiveness.[13]

The '80s and '90s also saw the development of multi-agent systems research. Craig Reynolds developed "boids" in 1986, showing that simple agent rules can produce global emergent behavior. This established the idea that societies of agents can produce intelligence exceeding the sum of their parts.[14] Contract Net Protocol, Knowledge Query and Manipulation Language, and early agent communication languages emerged during this period.

The ACT-R (2004) and Soar (2012) cognitive architectures are other examples of theoretical underpinnings that emerged or evolved during this period, integrating memory, reasoning and action toward multi-step cognition.

Concurrently, speech recognition improved dramatically. Chatbots and virtual assistants like Siri and Alexa became household names. But these systems were more like interfaces than independent actors. They could perform specific tasks, but they did not plan, reason or coordinate their behavior across multiple steps.

The prime agentic era (2010s-)

LLMs unlocked the potential of agents by giving them general-purpose reasoning capabilities. It's been a whirlwind of activity from that point forward.

Large language models

The most critical spark that ignited the revolution in agents that is currently underway did not emerge from classical agent research but from large language models (LLMs) equipped with natural language processing (NLP) capabilities. Starting with BERT in 2018 alongside OpenAI's GPT, LLMs became increasingly capable of general reasoning and communication.

These models were not themselves agents, but they provided a general-purpose cognitive core. For the first time, software could understand open-ended instructions and plan, possessing adaptability to handle new situations. Over time LLMs were seen as not just text generators and pattern matchers ("glorified autocomplete"), but as general problem-solvers.

Reinforcement learning and deeper learning

Another major machine learning technology that enabled the agentic era is reinforcement learning, which goes back at least as far as the 1980s. Before this technology, AI systems learned from fixed datasets. With reinforcement learning, agents can act in an environment, receive feedback and adjust their behavior to maximize a reward or minimize a punishment. This established the architecture of the classic agentic loop, which would later inspire modern agentic workflows. Reinforcement learning gave AI a mathematical language for agency, providing the formalism that enables a description of agentic behavior exhibited by physical robots or LLM-powered code-writing agents.

In 2017, reinforcement learning from human feedback (RLHF) emerged to improve alignment and reward modeling. In a 2017 paper, OpenAI's Paul F. Christiano, alongside other researchers from OpenAI and DeepMind, detailed RLHF's success in training AI models to perform intricate tasks like playing Atari games and simulating robotic locomotion.[15]

Before 2013, environments with high-dimensional inputs—like video games, robotics or real-world perception—were too complex for tabular or linear methods. But with newer deep learning methods, the technology became dominant. Google Deepmind unleashed the first version of the Deep Q-Network, a neural network that could learn to play Atari games directly from raw pixels and rewards. This was the first moment when an artificial agent combined representation learning with reinforcement learning in a fully end-to-end fashion. The decade saw more sophisticated approaches with more advanced algorithms such as AlphaGo, which combined deep neural networks with Monte Carlo Tree Search.

Learning robots

While deep reinforcement learning progressed in digital use cases like games and simulations, robotics researchers were advancing embodied agents that interacted with real-world environments—no longer pixels or abstractions, but objects, cameras, motors and physics. In the early 2010s, the PR2 robot—paired with the open-source Robot Operating System (ROS)—became the standard research platform for autonomous manipulation, navigation and multi-step task execution. These systems pioneered long-horizon action execution, task planning and multi-sensor integration.

The next major advancement came with RT-1 (Robotics Transformer), which used a large transformer model trained on demonstrations from hundreds of real-world robotics tasks. RT-1 was groundbreaking because it showed that multimodal foundation models could act as robotics agents, generalizing to new skills, objects and settings.

Prompt engineering

Early experiments in prompt engineering in the early 2020s hinted at emergent agentic behaviors, and the possibility of humans engaging with agents through natural language. Prompts like "first, think step-by-step, then answer the question" or "explain your reasoning, then provide your final answer," showed that LLMs could simulate procedural reasoning.

Chain-of-thought (CoT) prompting, in particular, revealed task decomposition, where an agent breaks down complex tasks into several smaller, more manageable operations. Later agentic frameworks would use steps like think, act, observe and repeat—a sequence later formalized in the ReAct framework. CoT was essentially a human-driven version of the agentic loop that would later be AI-driven.

Tool use

During this period researchers also introduced breakthroughs in autonomous tool calling and tool use, where agents would no longer require prompting to act. OpenAI's WebGPT demonstrated that an LLM could browse the internet, click links and retrieve information. Even more excitingly, LLMs with retrieval augmented generation (RAG) could merge outputs predicated on their own training data with new information sourced externally. What's more, augmenting agents with tool use allowed them to not only retrieve external information, but take actions in external environments.

Between 2022 and 2024, this new AI-powered cognitive capability enabled the agentic era in which we are now situated. Developers began to surround LLMs with all sorts of scaffolding, giving them access to APIs, knowledge bases and other software.

In 2023, Meta AI introduced Toolformer, the first LLM that could autonomously choose when and how to call external tools. This was the earliest demonstration that LLMs could generate and use their own action interfaces—a precursor to fully-fledged agent frameworks.

Later that year, OpenAI introduced function calling. For the first time, LLMs could call tools in a structured, predictable programmable format, pass parameters in JSON and interface with external systems through an explicit action space.

Multimodality

The development of large-scale multimodal generative AI models came in 2023-2024 with GPT-4 and GPT-4o. Now agents could "see" interfaces, interpret images, understand spatial layouts, converse in real time and integrate multimodal signals into planning.

Frameworks and protocols

Platforms and frameworks made building agents easier, even for users with limited software development skills. AutoGPT, BabyAGI, ReAct, AutoGen and AgentGPT demonstrated that LLMs could autonomously generate goals, strategically plan a course of action and learn from the outcomes of that strategy, all without human intervention. By streamlining a user experience with ready-made templates, LangChain became the first mainstream agent orchestration framework.

In late 2024, Anthropic introduced the Model Context Protocol (MCP), an open-source framework intended to standardize the way agents integrate and share data with external tools, systems and data sources. The following year, IBM introduced the Agent Communication Protocol (ACP), which aimed

to standardize the way agents communicate with one another. ACP later merged with Google's Agent2Agent (A2A) protocol under the Linux Foundation. Such AI agent protocols have helped facilitate a thriving ecosystem for AI agent development.

By 2025, "agentic AI' was the buzzword of the industry, and every player in the space was working on an agentic platform or solution of some form or another. Agents are now being rolled out in seemingly every industry, from supply chains to healthcare. It remains to be seen whether agents will live up to the massive amounts of hype surrounding them. We don't know if they will revolutionize life as we know it, but they remain one of the most promising technologies on the planet.