ALM | LAW.COM

New York Law Journal

TECHNOLOGY TODAY

ARTIFICIAL INTELLIGENCE

The Emerging World of AI Agents

By Katherine B. Forrest

April 30, 2024

If you are wondering what the next big thing in artificial intelligence (AI) is—what 2024 holds for us—I can tell you right now that it's AI agents. No doubt about it. Their capabilities are exciting, a little scary and essentially unknown. For legal practitioners, the issues that AI agents raise are complex.

There are a number of different definitions of AI agents (sometimes called AI assistants), but they share the following characteristics: a multimodal large language model (LLM) that can autonomously and sequentially carry out multidimensional goals, and that require it to move between domains and call on various tools or APIs to do so.

Let me break all of that techno-jargon down: AI agents are able—today—to take an instruction from say, a human, who asks the agent to carry out not one, but multiple tasks, by typing a prompt into an LLM. The agent can, by accessing tools on a computer to which it has been granted or can get permission (for instance, your bank account), autonomously proceed to take any number of steps across multiple digital environments, to accomplish what's been requested.

Katherine B. Forrest

An easy example of using an AI agent would be asking it to plan a vacation for you and your family, generally with certain monetary, date and location characteristics. The agent can then look through websites, find good deals, book tickets, book a hotel or Airbnb, book dinner reservations and then enter all of the details onto an itinerary for you on a single sheet of paper.

Another example is an agent asked to write code that is capable of self-testing, self-correcting and then in turn, instructing the code it has written to write more code.

Another example is an agent asked to interact with a geriatric patient socially but perceiving an issue, can make a doctor's appointment,

KATHERINE B. FORREST *is a partner in Paul, Weiss, Rifkind, Wharton & Garrison's litigation department and a member of the antitrust practice group. She previously served as a U.S. District Judge for the Southern District of New York and as a Deputy Assistant Attorney General in the U.S. Department of Justice's antitrust division.*

ADOBE STOCK

arrange for transportation, medicine delivery and the like.

This is a new world of AI opening up to us. AI agents can work for humans, functioning as our assistants, or on behalf of humans as our true agents; they can work singly or in AI agent groups, they can work under the direction of another AI agent or a human instructor.

This is not science fiction. If you want to read more about it—which I strongly suggest you and we all should—I refer you to DeepMind's extensive piece, published April 19, 2024, entitled "The Ethics of Advanced AI Assistants", a paper released by a group from Carnegie Mellon, Mathur, et al, entitled "Advancing Social Intelligence in AI Agents: Technical Challenges and Open Questions", published April 17, 2024, Sheng of N.Y.U.'s paper entitled "From Language Models to Practical Self-Improving AI Agents", published April 18, 2024, a paper by researchers from IBM and Microsoft (Masterman, Besen, Sawtell and Chao), entitled, "The Landscape of Emerging AI Agent Architectures for Reasoning, Planning and Tool Calling: A Survey", published April 17, 2024, among many others.

These papers share an acknowledgment of the significant capabilities of AI agents, and a need to align their reasoning, judgment and autonomous planning with human goals. Agents can enable complex use cases that make generative AI based on "one at a time" queries seem quaint. Agents that can be given multistep tasks that require reasoning, planning and problem solving, signify an exponential leap in what can be accomplished with these tools.

As we have seen already, AI tools of all kinds need to be monitored and steered toward human-aligned goals. Agents that can function autonomously, and with varying amounts of anonymity, imply new risk vectors. Guardrails need to be put into place to ensure that companies using agents have appropriate controls, and that companies exposed to agents are appropriately hardened to risks they can pose.

AI agents are able—today—to take an instruction from say, a human, who asks the agent to carry out not one, but multiple tasks, by typing a prompt into an LLM.

Let me give an example: to perform many tasks, AI agents need passwords. CAPTCHA is not a significant impediment. Once an agent has access to a password, if directed to divulge the password to a malicious actor, bad things could happen. One possible mitigation could be an expiration of permissions for passwords, or temporary passwords, and the like. You get my point.

Legal practitioners will be asked by their clients to assess risks associated with AI agents. Among those that should be flagged are ensuring that use of AI agents within a company is known and its permissions to access aspects of the internal systems carefully monitored. Cybersecurity teams will want to be sure to assess they can recognize requests to access internal systems that may be coming from external agents, and close off any risk vector found.

There will also be questions about using AI agents that may go rogue: doing more than requested or doing it differently than expected.

Liability questions may arise and questions about whether an AI Agent can act ultra vires, or whether insurance coverage is applicable. Appropriate diligence should be performed before any company licenses an AI agent.

Financial institutions will want to ensure that security protocols are updated to take into consideration that malicious actors may seek to have Agents fraudulently access accounts. There will also be new versions of age old tort claims where questions of proximate cause are raised and carefully analyzed.

With all of these protective measures should be taken into account, it is also true that agents will be the source of potential advances that will occur faster and more comprehensively than a human could accomplish. Imagine the ability to have a team of agents instructed to perform a series of tasks related to drug discovery and toxicity analysis; agents that can plan ahead a number of steps to stave off a problematic result.

We are only a few months into 2024, and we can see just the tip of what AI agents will be bringing our way. I, for one, am excited about their potential, wary of how they will be anthropomorphized, and wondering what we will be thinking about them in December. I'll report back on that, stay tuned…