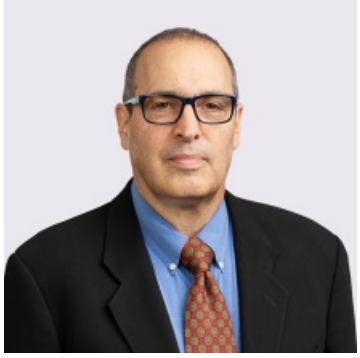
Legal Update Article

The Importance of Being Erroneous: Are AI Mistakes a Feature, Not a Bug?

By Thomas I. Barnett

February 19, 2025

Meet the Authors



Thomas I. Barnett

(Tom)

Principal and Chief Data Officer

Thomas.Barnett@jacksonlewis.com

Related Services

Artificial Intelligence & Automation

Takeaways

- Recent advances in autonomous AI agents could signal a breakthrough in how AI can eventually recognize and learn from mistakes.
- Some of our greatest innovators have recognized that making mistakes (and learning from them) is the key to innovation and invention.
- If AI agents can learn from mistakes and gain experience, they may be able to formulate and answer new questions.

Article

No one intentionally sets out to make a mistake. Yet, it has long been recognized by some of our greatest innovators from Thomas Edison to Albert Einstein to Henry Ford that making mistakes (and learning from them) is the key to innovation and invention. Some of history's most important discoveries started with mistakes: penicillin, radioactivity and X-ray technology, just to name a few. What we call a mistake may just be shorthand for our improvisational interaction with an unpredictable world. Or, as Oscar Wilde observed, "experience is simply the name we give our mistakes."Recent advances in autonomous AI agents could signal a breakthrough in how AI recognizes and learns from mistakes.

So, as society moves rapidly, if apprehensively, toward turning over more and more important tasks to AI, from menial labor to managing the electrical grid, it is worth considering how AI understands and grapples with mistakes and whether it meaningfully differs from the human approach. Answering that question could prove more significant than analyzing AI successes.

AI companies spend a lot of time telling us they are doing their best to eliminate mistakes. That's reassuring when it comes to air traffic control and cancer diagnosis. But what if incorporating mistakes into our thinking is actually *the most important ingredient* in advancing our knowledge of the world? What if looking at an error and then refining our definition of the problem is the key to how we advance our thinking rather than simply rejecting the mistake outright and moving on? Einstein was well known for emphasizing the significance of defining a problem over its solution; given one hour to save the planet, he would note that he would spend 59 minutes defining the problem and one minute resolving it.

When you break it down, computers operate in a completely binary universe — on/off, 1/0, black or white, with no nuance in between. For purely mathematical or empirical problems that works well. Computers are much faster at solving math problems than we are, but humans see shades of gray. We know not every problem has a simple yes or no answer. Indeed, as the level of abstraction increases with real-world problems, what is

right or wrong often becomes murkier and more subjective. You can program a self-driving car to swerve away from an object in the road but what if the object is a cardboard box and swerving would result in hitting a crowd of pedestrians on the sidewalk? That's where *human* judgment comes in — something we apply instinctively every day but something which we have yet to reduce to computer code.

Originating in the 1950s, "machine learning" has used algorithms to analyze data and compare it to previous human judgments (*supervised* machine learning) or identifying patterns in data (*unsupervised* machine learning). But machine learning, even as it has advanced, is not designed to look at a mistake, then, in response, completely redefine the problem in a way that transforms the incorrect result into a useful outcome for the newly redefined problem, like when unintended mold growing on a Petri dish in a messy lab in London led to the inadvertent discovery of penicillin. Of course, not all mistakes are created equal — sometimes a mistake is just a mistake, not a hidden goldmine. Machine learning algorithms so far only take incorrect answers and try to use them to avoid making the same error the next time — not rethink the whole problem.

Generative AI tools have been designed to function by being pretrained (the *P* in GPT) on vast amounts of data such as large language models or other data sources and then to provide responses to inputs or prompts (a question or instruction) provided by users. A new approach allows AI to interact directly and more autonomously with data and react in a dynamic way — a lot more like what humans do. This relies on what are referred to as *AI agents,* which Bill Gates wrote are "going to upend the software industry, bringing about the biggest revolution in computing since we went from typing commands to tapping on icons." That may be an understatement. AI agents are now being designed to make decisions without human intervention to perform predefined (for now) tasks. It can reach into the outside world, find data it hadn't previously encountered, analyze it, then take action — more like humans and less relying on the fixed data universe of a chess program or a chatbot that cannot go beyond its pretrained knowledge.

AI agents learn from mistakes, gain experience and may eventually use them to formulate and answer new questions. For as Edison observed, "I have not failed, I just found 10,000 ways that won't work."

Please contact the author with any questions.