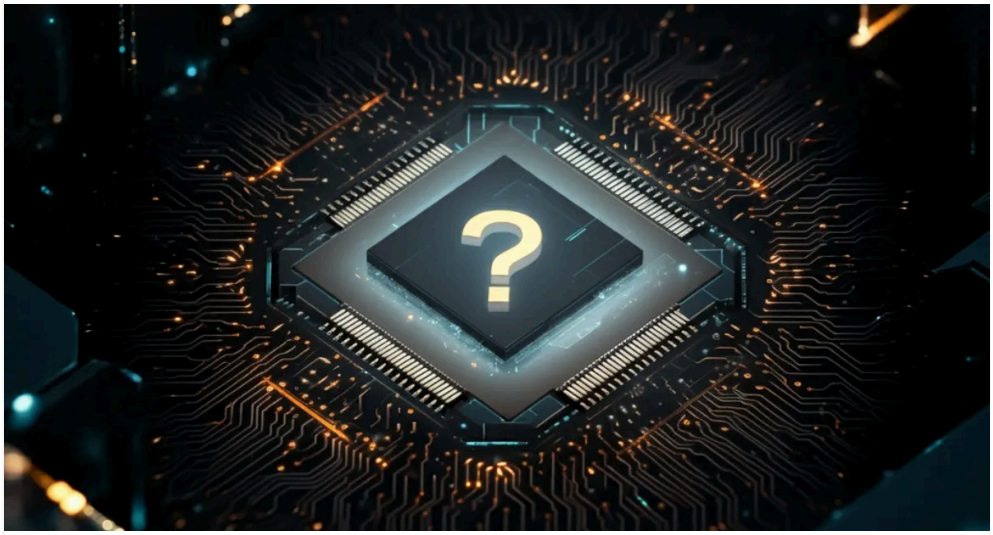
AI'S NOT TO REASON WHY (BECAUSE WE DON'T KNOW IF IT CAN)

Thom Hawkins, Erin Zaroukian and Adrienne Raglin | 11.01.24



When a unit receives an order to engage a target at a specified time and location, many questions must be answered. Can the unit physically get to that location at that time? Do its members have the right weapons and other equipment? What about the skills necessary to engage the target? Each of these inquiries raises more questions—about terrain and vehicles, ammunition and fuel supplies, training and aptitude of individual soldiers.

The thought process used to determine whether the unit is mission capable is based on logical inference. To move forward, a vehicle needs an operator and fuel, for example. The amount of fuel it needs depends on a number of factors, including distance and terrain. The time it takes to traverse a route is based on distance and vehicle speed over the terrain. To answer the question, then, of whether the unit can complete this mission, we need to know many different factors and how they are related. We can use a structured data model such as an ontology to represent what we know and to facilitate answering specific questions with certainty. However, a comprehensive ontology can be resource intensive to develop and can be challenging to adapt to dynamic environments, where not only the data may change, but also the relationships between entities.

Early AI systems used an *expert approach* (i.e., modeling expert judgments for a given field). Developers built these systems with binary decision points, prescribing right or wrong answers. That is, the AI was certain, at least with respect to the worldview of the expert on which the system was based. The expert approach worked on a limited basis, but it did not scale well, nor did it handle the uncertainty inherent in the fog of war. It's not possible to create a perfect model of the world, and any attempt to do so risks oversimplifying a given scenario. Our ability to model and predict weather has repeatedly demonstrated its shortfalls, for example.

Our understanding of how humans reason and make decisions continues to evolve, but has moved past the idea that we do so with certainty. Experienced leaders make decisions not only based on external factors, but also consider their own intuition, developed through experience in the field. The limitations of expert AI systems led to an AI winter starting in the 1990s. Now, new AI tools such as large language

SEARCH …

FOLLOW US

FACEBOOK

YOUTUBE

TWITTER

DISCLAIMER

The articles and other content which appear on the Modern War Institute website are unofficial expressions of opinion. The views expressed are those of the authors, and do not reflect the official position of the United States Military Academy, Department of the Army, or Department of Defense.

The Modern War Institute does not screen articles to fit a particular editorial agenda, nor endorse or advocate material that is published. Rather, the Modern War Institute provides a forum for professionals to share opinions and cultivate ideas. Comments will be moderated before posting to ensure logical, professional, and courteous application to article content.

UPCOMING EVENTS

There are no upcoming events.

ANNOUNCEMENTS

Announcing the Modern War Institute's 2024–2025 Research Fellows

We're Looking for Officers to Join the Modern War Institute and the Defens...

Call for Applications: MWI's 2024–25 Research Fellows Program

Join Us Friday, April 26 for a Livestream of the 2024 Hagel Lecture, Featuring...

models, or LLMs, are reopening up the conversation of how intelligent systems could support thinking and reasoning for planning tasks.

LLMs Have Entered the Chat

Over the past two years, ChatGPT and similar LLMs have both impressed and disappointed their users. Their increasing ability to mimic written language has made them go-to platforms for students and content developers, but at the same time, we're eager to mock nonsensical responses to straightforward questions. After the initial hype, most people have not moved past using the technology for social media attention, while a small subset of die-hard users have learned how to use the technology in less trivial ways, such as real-time language translation. Interest among advocates within the military community has driven the development of clones that make use of the same open-source models as ChatGPT and Meta's Llama but are placed behind a firewall to avoid leaking sensitive information.

Besides enterprise tasks, LLMs are also being considered for integration, as an enabler, into tactical systems. An LLM can be used, for example, to provide a natural language interface, translate languages, expedite the production of operations orders.

However, in contrast to expert AI systems, LLMs are all about probabilities (e.g., given a corpus of text, what word will be most likely to follow another), which precludes the possibility of reasoning to a single answer or solution. A funny thing happened, though, when companies like OpenAI began to scale their models—more text, more parameters, more compute. The LLMs seemed to develop new abilities.

In his 1972 paper, "More Is Different," P. W. Anderson that came to define the concept of emergence for LLM researchers. Emergence occurs "when quantitative changes in a system result in qualitative changes in behavior." In terms of LLMs, we define an ability like reasoning as emergent "if it is not present in smaller models but is present in larger models." There is, however, no magic number that represents a threshold for emergence. In addition, once we recognize that an ability emerged, we may find creative ways to access it even with smaller models.

Whether LLMs demonstrate an emergent capability for reasoning is debatable. One study found that, when confronted with a novel logic problem, LLMs showed evidence of a nascent reasoning ability, though when solving well-known logic problems, LLMs tended to produce the response most frequently associated with the puzzle, even if it did not fit that particular formulation, casting doubt on whether the model truly understands the problem. Other research showed that an LLM could be confused by details a human would recognize as irrelevant to the question asked.

Another consideration is that just because an LLM can rationalize an argument does not mean that the conclusion is correct. For example, if we state that all birds can fly and that all penguins are birds, it is rational to say that penguins can fly. The fact that they cannot is not an issue with the logic—it is that the first premise, all birds can fly, is not true. (One could also swap out penguin for Osprey in this argument.)

Machine Psychology

Thomas Sheridan, in his book *Telerobotics, Automation, and Human Supervisory Control*, identified several "trust-causation factors" that would build trust between a human and a machine, including reliability and understandability. One of the features of a probabilistic system like an LLM is that a user will not always get the same answer to the same question, which presents a challenge for test and evaluation of our systems, as well as for reliance on the output. The sensitivity of the input to the output could be controlled by using detailed prompts, or even building prompts into our systems to ensure the response falls within an expected range.

One of the recent innovations in LLMs is chain-of-thought prompting, which encourages an LLM to eat the elephant one bite at a time by breaking a problem down into its constituent parts rather than using its predictive abilities to divine the likely answer from its model. (OpenAI recently released a model that does this specifically.)

Chain-of-thought prompting has already aided the fledgling field of machine psychology, which studies the behavior of LLMs using experiments inspired by those used to study reasoning in humans. For example, researchers have seen human-like reasoning biases emerge as models grow (e.g., reasoning incorrectly that if a bat and ball together cost $1.10 and the bat costs $1.00 more than the ball, then the ball must cost $0.10), but these biases abate both in sufficiently large models *and* in smaller models *if* they are prompted to use chain-of-thought reasoning.

Because LLMs are built by training them on a largely uncurated corpus of text, many of our own biases and ways of thinking and reasoning are included in the models. How we reason is based on the evolution of a specific kind of intelligence that humans have found necessary for survival in our environment and social situations. Human intelligence is not necessarily universal. If we lived on another planet, our intelligence would have evolved in response to a different environment or different problems.

Nearly seventy-five years ago, Paul Fitts published what has since become known as the Fitts list, an enumeration of skills that the humans are better at and those that machines are better at. While it stands to reason that with all the advances since 1951 the list would be outmoded, it is still a largely valid framework. Researchers still consult Fitts's list and develop machines to improve on the skills that humans do. LLMs or related technology may eventually be as good or better than humans at reasoning, but the utility will be limited if it merely replaces, rather than supplements, our own judgment.

At an Army workshop regarding AI-generated courses of action, an officer expressed his concern about presenting a recommended course of action to his commander without being able to articulate *why* it was recommended. If prompted to do so, an LLM will explain how it arrived at an answer. However, we do not yet know whether this explanation is reliable. It could be another hallucination, just as an LLM will make up facts to ensure it provides the user an acceptable response from the perspective of context, without a notion of what is true or real.

This leaves us with three alternatives: (1) field this technology and trust that our soldiers will show discretion in how they employ it; (2) bar its use until we can improve the reliability to acceptable levels, just as we would with any system we field; and (3) educate its users on the constraints and limitations of LLMs as reasoning and research tools. These options are not necessarily mutually exclusive. One path forward that includes all three is human-machine teaming, rather than delegation.

Much, perhaps most, of the current public discussions about leveraging AI capabilities center on a model that pairs humans with AI tools. But the argument in favor of human-machine teaming is most often made from a perspective of maximizing capability by combining what we do best with what an AI system does best. The problem is not only about maximizing capability, however; it is also about overcoming the limitations of AI tools like LLMs and mitigating the risk that arises from the fact that we don't always know what happens inside AI's black box. AI may help us think faster, but it is not ready to replace our thinking.

Thom Hawkins is a senior information scientist with US Army Project Manager Mission Command and a PhD student at the Naval Postgraduate School where he studies teaming and collaboration between humans and near-peer AI.

Dr. Erin Zaroukian is a cognitive scientist with the DEVCOM Army Research Laboratory studying artificial reasoning and human-robot collaboration.

Dr. Adrienne Raglin is an electronics engineer with the DEVCOM Army Research Laboratory researching artificial reasoning and decision making.

The views expressed are those of the authors and do not reflect the official position of the United States Military Academy, Department of the Army, or Department of Defense.

© 2025 Modern War Institute