EXPERT VOICES

AI's challenge of understanding the world

MELANIE MITCHELL, Authors Info & Affiliations

SCIENCE 10 Nov 2023 Vol 382, Issue 6671 DOI: 10.1126/science.adm8175

In thinking about the challenge of getting artificial intelligence (AI) to understand our complex world, I recalled a Twitter post from a user of Tesla's self-driving system. The user tweeted that his car kept stopping abruptly at a particular location for no apparent reason. Then he noticed a billboard advertisement on the side of the road, featuring a sheriff holding up a stop sign. The car's vision system had interpreted this as an actual stop sign, and slammed on the brakes.

Such failures of understanding—recognizing a stop sign but not its crucial context—are common in AI applications. Computer vision networks can fail when objects appear in unusual conditions; language translation software can misinterpret meanings in high-stakes situations; and medical diagnosis systems can misconstrue what they should learn from the data they are trained on. If the AI systems rapidly making their way into every corner of our lives are to become more useful, trustworthy, transparent, and safe, then they need to achieve a deeper commonsense understanding of our world.

Some AI leaders have declared that the remarkable capabilities of large language models (LLMs) and other "generative" AI systems have finally crossed a barrier of understanding, and that we are already seeing the arrival of humanlike AI. After all, these systems exhibit uncanny abilities to converse with us in natural language, generate realistic images in response to our prompts, write functioning computer code, and even excel on standardized exams designed to test human reasoning abilities.

Nevertheless, the question of how much LLMs really understand about the world is the subject of a polarized debate. Scholars on the "no understanding" side predict that machines trained only on language "will never approximate human intelligence, even if trained from now until the heat death of the universe." Other researchers asse that the behavior of LLMs arises not from grasping the meaning of language but rather from learning complex p terns of statistical associations among words and phrases in training data and later performing "approximate retrieval" of these patterns and applying them to new queries.

Indeed, several recent studies have cast doubt on the robustness of LLMs' abilities for generalization and abstraction, showing that these systems are unreliable at solving problems or dealing with situations that are sufficiently different from those that appeared in their training data. The propensity of LLMs for "hallucinating" answers to queries and their vulnerability to adversarial attacks indicate a lack of grounding in the real world, including the intents behind their users' requests.

Current AI systems seem to be lacking a crucial aspect of human intelligence: rich internal models of the world. A tenet of modern cognitive science is that humans are not simply conditioned-reflex machines; instead, we have inside our heads abstracted models of the physical and social worlds that reflect the causes of events rather than merely correlations among them. We rely on these mental models to simulate and predict the likely results of possible actions, to reason and plan in unfamiliar situations, to imagine counterfactuals ("what would have happened if I hadn't stopped the car in time?"), and to update our knowledge and beliefs on the basis of experiences. Moreover, we have mental models not only of the external world and other people, but of ourselves, enabling us to assess and explain our reasoning and decision-making processes. How such models are implemented in our brains is much debated, but there is little doubt that they are foundational to our intelligence.

The problem of acquiring "world models" has been a focus of AI research for many decades. Researchers have experimented with many methods for either manually programming such models or for trying to get machines to learn them from data or experience. These efforts have had some success in AI domains with simplified "worlds," such as playing video games and in robot control tasks.

However, LLMs and other generative AI systems are a different ballgame. No one has programmed any world models, nor are these systems explicitly trained to learn them. Instead, generative AI systems are typically trained with sequences of "tokens"—parts of words or images—and are asked to predict the next token in the sequence. Yet, these enormous models, after being trained on trillions of tokens taken from digitized text and images, seem to have grasped some basic aspects of the world and of human society. Is it possible that humanlike world models have emerged in these systems, even though they were never explicitly programmed or learned? This is precisely what some in the AI community are claiming.

For example, in a recent interview, Ilya Sutskever, cofounder and chief scientist at OpenAI, said:

When we train a large neural network to accurately predict the next word in lots of different texts…it is learning a world model.… This text is actually a projection of the world.… What the neural network is learning is more and more aspects of the world, of people, of the human conditions, their hopes, dreams, and motivations…the neural network learns a compressed, abstract, usable representation of that.

This is a provocative hypothesis about LLMs—but what evidence is there to back it up?

One recent study explored whether a language model implicitly learns a "world model" in the context of the board game Othello, which is played on an eight square–by–eight square board. A game can be described by listing the sequence of moves, with the positions on the board labeled by row (letter) and column (number). For example, a sequence could start with player 1 placing a black piece on square F5, followed by player 2 placing a white piece on square F6, and so on.

The researchers used an Othello game simulator to generate 20 million sequences like this, each consisting of a partial game. There was no expertise or strategy involved; each item in the sequence was a randomly chosen legal move, derived from the moves that appeared before. These sequences were then used to train a neural network (an eight-layer "transformer" model) called OthelloGPT. The neural network had no knowledge of the game's rules or even that the input sequences represented a game. All that it saw were sequences of text tokens (e.g., token F5 followed by token F6, etc.). Analogous to LLMs trained on natural language, OthelloGPT was trained to predict wh of 64 possible tokens would appear next in the sequence.

After its training, OthelloGPT could accurately predict legal moves, even on sequences that it had never seen in its training data. How did it do this ? Was it relying on statistical correlations among patterns of tokens in its training

sequences, or had it learned, as Ilya Sutskever said, a compressed, abstract, usable model of the "world"—the board, the pieces, the players, and the rules of the game?

To address this question, the researchers used "probes" to determine what OthelloGPT had learned. A probe is a simpler neural network that is trained to decode the original neural network's internal activations—the "firing" activity of simulated neurons in the network's internal layers in response to an input. The researchers trained probes to predict—using only activations in various layers of OthelloGPT—whether, after a particular sequence of moves, a given square contained a black or white piece, or no piece at all. Even though OthelloGPT had been trained only on sequences of text tokens, its internal activations could be decoded to predict what pieces were in what positions at particular times in a game.

Moreover, the researchers showed, through clever manipulations of OthelloGPT's internal activations, that it was not just encoding the state of the board as a side effect but was using this internal representation— a "world model"— to predict legal moves.

This result is a fascinating proof of principle: Nontrivial, useful internal representations of a simple "world" can emerge from language-model training. Other research groups found analogous results for language models implicitly encoding concepts of color spaces, spatial direction, and world states of simple text-adventure games.

However, there is a chasm between these results on ultrasimple "worlds" and Ilya Sutskever's contention that ChatGPT has learned—from trillions of sequences of text tokens—immensely complex, actionable models of the real world and its human inhabitants. Even for the simple Othello example, a humanlike world model would do much more than encode the state of the board; it would encode the rules of the game, enable reasoning about gameplay strategies, and equip the system to respond to moves that are quite different from its training experiences and even to adapt flexibly to new variations of the game. Moreover, such a world model would help the system explain its knowledge and decision-making to others. Such general abilities are hallmarks of human understanding, but despite remarkable progress, current AI systems have not yet caught up.

It's an open question whether current machine-learning paradigms will yield the kinds of understanding needed for trustworthy AI in the real world, or whether new kinds of paradigms are needed, such as combining language models with symbolic approaches, incorporating new ideas from reinforcement learning, creating integrated cognitive architectures, or including embodied experience. To trust the AI systems that will inevitably be ubiquitous in our world, we face twin challenges: first, enabling these systems to usefully understand that world, and second, equipping ourselves with the scientific tools to make sense of how they do it.

eLetters (1)

eLetters is a forum for ongoing peer review. eLetters are not edited, proofread, or indexed, but they are screened. eLetters should provide substantive and scholarly commentary on the article. Embedded figures cannot be submitted, and we discourage the use of figures within eLetters in general. If a figure is essential, please include a link to the figure within the text of the eLetter. Please read our Terms of Service before submitting an eLetter.

LOG IN TO SUBMIT A RESPONSE

NOV. 9, 2023

Living with AI, Intelligently

BARRY J MCKENNA Independent researcher, none

Full Text

Help

AI and LLMs are a class of technological product for which billions of human-machine dollars have been invested. The qu... tion of intelligence and machines, however, begs the question of how we might replicate the emodied experience of the evolution of life's intelligence and billions of embodied cells - the foundation for the essence of mind and intelligent survival...

view more

Integrated platform for multiscale molecular imaging and phenotyping of the human brain

BY JUHYUK PARK, JI WANG, *ET AL.*

Metabolic inflexibility promotes mitochondrial health during liver regeneration

BY XUN WANG, CAMERON J. MENEZES, *ET AL.*

Artemisinins ameliorate polycystic ovarian syndrome by mediating LONP1-CYP11A1 interaction

BY YANG LIU, JING-JING JIANG, *ET AL.*

Full Text

TABLE OF CONTENTS

Help

LATEST NEWS

SCIENCEINSIDER | 14 JUN 2024

As U.K. election nears, major parties reveal their science policies

SCIENCEINSIDER | 14 JUN 2024

Amid Russian attacks, Ukrainian astronomers fight to salvage a unique observatory

NEWS | 13 JUN 2024

News at a glance: Ancient malaria, geoengineering controversy, and science fraud law

NEWS FEATURE | 13 JUN 2024

App developed with Indigenous trackers helps almost anyone monitor wildlife

SCIENCEINSIDER | 13 JUN 2024

Wild poliovirus spreads across Pakistan and Afghanistan

SCIENCEINSIDER | 12 JUN 2024

Record settlement over China funding puts U.S. research institutions on notice

RELATED JOBS

RECOMMENDED

12 JUL 2019 | BY KATIE CAMERO

Artificial intelligence conquers world's most complex poker game

SPECIAL NEWS REPORT | SEPTEMBER 1997

'RoboCup' Soccer Match Is a Challenge for Silicon Rookies

15 AUG 2018 | BY MATTHEW HUTSON

To hone its collaborative skills, this AI is taking on the world's top video game players

RESEARCH ARTICLE | JANUARY 2018

Superhuman AI for heads-up no-limit poker: Libratus beats top professionals

View full text

About News from Science

COMMENTARY

Opinion

Analysis

Blogs

JOURNALS

Science

Science Advances

Science Immunology

Science Robotics

Science Signaling

Science Translational Medicine

Science Partner Journals

AUTHORS & REVIEWERS

Information for Authors

Information for Reviewers

LIBRARIANS

Manage Your Institutional Subscription

Library Admin Portal

Request a Quote

Librarian FAQs

ADVERTISERS

Advertising Kits

Custom Publishing Info

Post a Job

RELATED SITES

AAAS.org

AAAS Communities

EurekAlert!

Science in the Classroom

ABOUT US

Leadership

Work at AAAS

Prizes and Awards

HELP

FAQs

Access and Subscriptions

Order a Single Issue

Reprints and Permissions

TOC Alerts and RSS Feeds

Contact Us