EXPERT VOICES

Debates on the nature of artificial general intelligence

MELANIE MITCHELL Authors Info & Affiliations

SCIENCE 21 Mar 2024 Vol 383, Issue 6689 DOI: 10.1126/science.ado7069

The term "artificial general intelligence" (AGI) has become ubiquitous in current discourse around AI. OpenAI states that its mission is "to ensure that artificial general intelligence benefits all of humanity." DeepMind's company vision statement notes that "artificial general intelligence…has the potential to drive one of the greatest transformations in history." AGI is mentioned prominently in the UK government's National AI Strategy and in US government AI documents. Microsoft researchers recently claimed evidence of "sparks of AGI" in the large language model GPT-4, and current and former Google executives proclaimed that "AGI is already here." The question of whether GPT-4 is an "AGI algorithm" is at the center of a lawsuit filed by Elon Musk against OpenAI.

Given the pervasiveness of AGI talk in business, government, and the media, one could not be blamed for assuming that the meaning of the term is established and agreed upon. However, the opposite is true: What AGI means, or whether it means anything coherent at all, is hotly debated in the AI community. And the meaning and likely consequences of AGI have become more than just an academic dispute over an arcane term. The world's biggest tech companies and entire governments are making important decisions on the basis of what they think AGI will entail. But a deep dive into speculations about AGI reveals that many AI practitioners have starkly different views on the nature of intelligence than do those who study human and animal cognition—differences that matter for understanding the present and predicting the likely future of machine intelligence.

The original goal of the AI field was to create machines with general intelligence comparable to that of humans. Early AI pioneers were optimistic: In 1965, Herbert Simon predicted in his book *The Shape of Automation for Men and Management* that "machines will be capable, within twenty years, of doing any work that a man can do," and, in a 1970 issue of *Life* magazine, Marvin Minsky is quoted as declaring that, "In from three to eight years we will have a machine with the general intelligence of an average human being. I mean a machine that will be able to read Shakespeare, grease a car, play office politics, tell a joke, have a fight."

These sanguine predictions did not come to pass. In the following decades, the only successful AI systems were narrow rather than general—they could perform only a single task or a limited scope of tasks (e.g., the speech recognition software on your phone can transcribe your dictation but cannot intelligently respond to it). The term "AGI" was coined in the early 2000s to recapture the original lofty aspirations of AI pioneers, seeking a renewed focus on "attempts to study and reproduce intelligence as a whole in a domain independent way."

This pursuit remained a rather obscure corner of the AI landscape until quite recently, when leading AI companies pinpointed the achievement of AGI as their primary goal, and noted AI "doomers" declared the existential threat from AGI as their number one fear. Many AI practitioners have speculated on the timeline to AGI, one predicting, for example, "a 50% chance that we have AGI by 2028." Others question the very premise of AGI, calling it vague and ill-defined; one prominent researcher tweeted that "The whole concept is unscientific, and people should be embarrassed to even use the term."

Whereas early AGI proponents believed that machines would soon take on all human activities, researchers have learned the hard way that creating AI systems that can beat you at chess or answer your search queries is a lot easier than building a robot to fold your laundry or fix your plumbing. The definition of AGI was adjusted accordingly to include only so-called "cognitive tasks." DeepMind cofounder Demis Hassabis defines AGI as a system that "should be able to do pretty much any cognitive task that humans can do," and OpenAI describes it as "highly autonomous systems that outperform humans at most economically valuable work," where "most" leaves out tasks requiring the physical intelligence that will likely elude robots for some time.

The notion of "intelligence" in AI—cognitive or otherwise—is often framed in terms of an individual agent optimizing for a reward or goal. One influential paper defined general intelligence as "an agent's ability to achieve goals in a wide range of environments"; another stated that "intelligence, and its associated abilities, can be understood as subserving the maximisation of reward." Indeed, this is how current-day AI works—the computer program AlphaGo, for example, is trained to optimize a particular reward function ("win the game"), and GPT-4 is trained to optimize another kind of reward function ("predict the next word in a phrase").

This view of intelligence leads to another speculation held by some AI researchers: Once an AI system achieves AGI, it will quickly achieve superhuman intelligence by applying its optimization power to its own software, recursively advancing its own intelligence and quickly becoming, in one extreme prediction, "thousands or millions of times more intelligent than we are."

This focus on optimization has led some in the AI community to worry about the existential risk to humanity from "unaligned" AGI that diverges, maybe crazily, from its creator's goals. In his 2014 book *Superintelligence*, philosopher Nick Bostrom proposed a now-famous thought experiment: He imagined humans giving a superintelligent AI system the goal of optimizing the production of paper clips. Taking this goal quite literally, the AI system then uses its genius to gain control over all Earth's resources and transforms everything into paper clips. Of course, the humans did not intend the destruction of Earth and humanity to make more paper clips, but they neglected to mention that in the instructions. AI researcher Yoshua Bengio provides his own thought experiment: "[W]e may ask an AI to fix climate change and it may design a virus that decimates the human population because our instructions were not clear enough on what harm meant and humans are actually the main obstacle to fixing the climate crisis."

Such speculative views of AGI (and "superintelligence") differ from views held by people who study biological intelligence, especially human cognition. Whereas cognitive science has no rigorous definition of "general intelligence" or consensus on the extent to which humans, or any type of system, can have it, most cognitive scientists would agree that intelligence is not a quantity that can be measured on a single scale and arbitrarily dialed up and down but rather a complex integration of general and specialized capabilities that are, for the most part, adaptive in a specific evolutionary niche.

Many who study biological intelligence are also skeptical that so-called "cognitive" aspects of intelligence can be separated from its other modes and captured in a disembodied machine. Psychologists have shown that important aspects of human intelligence are grounded in one's embodied physical and emotional experiences. Evidence also shows that individual intelligence is deeply reliant on one's participation in social and cultural environments. The abilities to understand, coordinate with, and learn from other people are likely much more important to a person's success in accomplishing goals than is an individual's "optimization power."

Moreover, unlike the hypothetical paper clip–maximizing AI, human intelligence is not centered on the optimization of fixed goals; instead, a person's goals are formed through complex integration of innate needs and the social and cultural environment that supports their intelligence. And unlike the superin-

telligent paper clip maximizer, increased intelligence is precisely what enables us to have better insight into other people's intentions as well as the likely effects of our own actions, and to modify those actions accordingly. As the philosopher Katja Grace writes, "The idea of taking over the universe as a substep is entirely laughable for almost any human goal. So why do we think that AI goals are different?"

The specter of a machine improving its own software to increase its intelligence by orders of magnitude also diverges from the biological view of intelligence as a highly complex system that goes beyond an isolated brain. If human-level intelligence requires a complex integration of different cognitive capabilities as well as a scaffolding in society and culture, it is likely that the "intelligence" level of a system will not have seamless access to the "software" level, just as we humans cannot easily engineer our brains (or our genes) to make ourselves smarter. However, we as a collective have increased our effective intelligence through external technological tools, such as computers, and by building cultural institutions, such as schools, libraries, and the internet.

What AGI means and whether it is a coherent concept are still under debate. Moreover, speculations about what AGI machines will be able to do are largely based on intuitions rather than scientific evidence. But how much can such intuitions be trusted? The history of AI has repeatedly disproved our intuitions about intelligence. Many early AI pioneers thought that machines programmed with logic would capture the full spectrum of human intelligence. Other scholars predicted that getting a machine to beat humans at chess, or to translate between languages, or to hold a conversation, would require it to have general human-level intelligence, only to be proven wrong. At each step in the evolution of AI, human-level intelligence turned out to be more complex than researchers expected. Will current speculations about machine intelligence prove similarly wrongheaded? And could we develop a more rigorous and general science of intelligence to answer such questions?

It is not clear whether a science of AI would be more like the science of human intelligence or more like, say, astrobiology, which makes predictions about what life might be like on other planets. Making predictions about something that has never been seen and might not even exist, whether that is extraterrestrial life or superintelligent machines, will require theories grounded in general principles. In the end, the meaning and consequences of "AGI" will not be settled by debates in the media, lawsuits, or our intuitions and speculations but by long-term scientific investigation of such principles.

eLetters (0)

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

No eLetters have been published for this article yet.

What We Should Not Forget about Down Syndrome
D.C. Fragoso et al., Neurographics, 2021

Powered by TREND MD

CURRENT ISSUE

Apoptotic cell identity induces distinct functional responses to IL-4 in efferocytic macrophages

BY IMKE LIEBOLD, AMIRAH AL JAWAZNEH, *ET AL.*

Removal of *Pseudomonas* type IV pili by a small RNA virus

BY JIRAPAT THONGCHOL, ZIHAO YU, *ET AL.*

Assessing the health burden from air pollution

BY TORBEN SIGSGAARD, BARBARA HOFFMANN

TABLE OF CONTENTS >

LATEST NEWS

SCIENCEINSIDER | 8 APR 2024

Efforts to support Palestinian scientists struggle with the realities of war

NEWS | 5 APR 2024

Insect poetry, conquering rats, and more stories you might have missed this week

NEWS | 5 APR 2024

Clearer skies may be accelerating global warming

SCIENCEINSIDER | 5 APR 2024

NSF tests ways to improve research security without disrupting peer review

SCIENCEINSIDER | 5 APR 2024

Australian museum's plan to cut research draws fire from scientists

SCIENCEINSIDER | 4 APR 2024

Bird flu may be spreading in cows via milking and herd transport

RELATED JOBS

Research Scientist - GI Med Oncology

University of Texas MD Anderson Cancer Center
Houston, Texas

Senior Research Assistant - Systems Biology (Spatial Transcriptomics)

University of Texas MD Anderson Cancer Center
Houston, Texas

Research Assistant II - Epigenetics & Molecular Carcinogenesis (Ishak Laboratory)

University of Texas MD Anderson Cancer Center
Houston, Texas

MORE JOBS ▶

RECOMMENDED

3 MAR 2017 | BY TONYA RILEY

Artificial intelligence goes deep to beat humans at poker

RESEARCH ARTICLE | MAY 2017

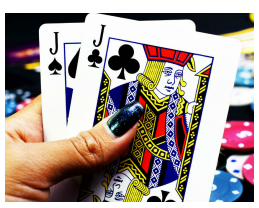
DeepStack: Expert-level artificial intelligence in heads-up no-limit poker

PERSPECTIVE | JULY 2023

Improving artificial intelligence with games

30 MAY 2019 | BY EDD GENT

Artificial intelligence learns teamwork in a deadly game of capture the flag



12 JUL 2019 | BY KATIE CAMERO

Artificial intelligence conquers world's most complex poker game

View full text