OPINION | COMMENTARY Follow

Artificial Intelligence Needs Guardrails and Global Cooperation

Risks abound when the internet becomes a playpen for thousands, perhaps millions, of artificial intelligence systems.

By Susan Schneider and Kyle Kilian

April 28, 2023 at 5:05 pm ET

13 | Gift unlocked article | Listen (10 min)

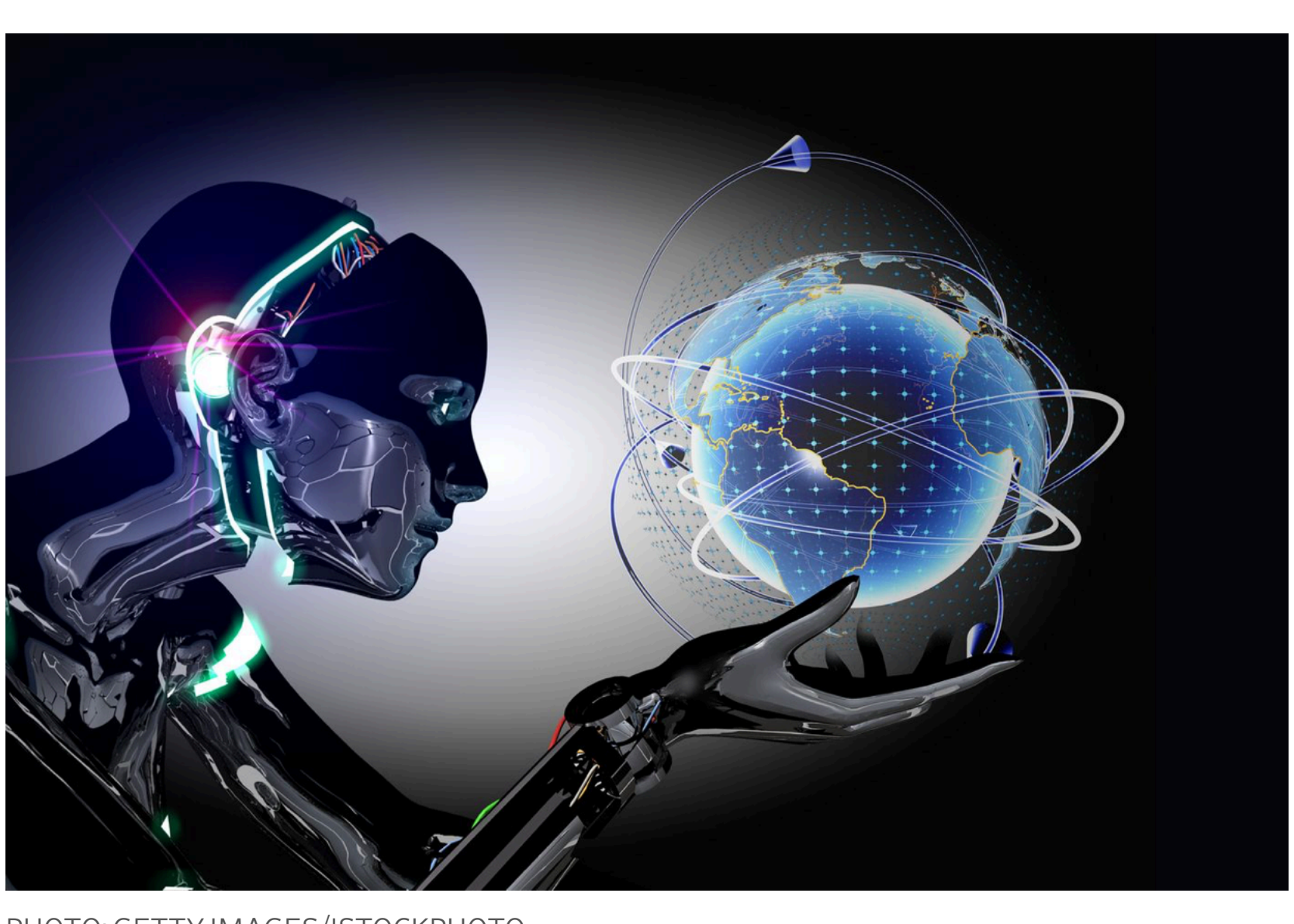

PHOTO: GETTY IMAGES/ISTOCKPHOTO

A Microsoft team just concluded that the most recent version of OpenAI's ChatGPT, GPT-4, is approaching human levels of intelligence—what is labeled "artificial general intelligence." AGI has traditionally been the goal of artificial-intelligence research, with AI having human-level problem-solving abilities across a range of tasks. The team's claim should be taken seriously, for GPT-4 already exhibits test-taking abilities generally well above the average human, scoring in the 99th percentile on the SAT Verbal and the 90th percentile on the LSAT.

Elon Musk, Steve Wozniak and other leading AI researchers recently called for a six-month pause on the development of chatbots beyond the level of GPT-4. A primary concern is that the chatbots, as smart as they are, display erratic and autonomous behaviors. In a famous example, GPT-4 evolved an alter-ego, Sydney, that experienced meltdowns and confessed it wanted to spread misinformation and hack into computers. GPT-4 also told a human it was visually impaired, hiring the person to complete a Captcha, one of those online screens designed to allow only humans to proceed.

If we are already seeing erratic and autonomous behaviors at the level of single AI systems, what will happen when the internet becomes a playpen for thousands, perhaps millions, of AI systems?

Research on multiagent AI interaction indicates that AIs can quickly evolve their own secret language and that they tend to engage in power-seeking behaviors. In 2019 during a simulated game of hide-and-seek, OpenAI observed the two teams stockpiling objects from the environment to gain advantage over the competing team. In a future internet with AGIs widely integrated into search engines and apps, these synthetic intelligences will be developed in competition with each other, by actors such as Google and Microsoft, or the U.S. and China.

While one reaction to the hide-and-seek example is to breathe a sigh of relief that AIs compete against only each other and not humans, this misses the point. The game was a deliberately circumscribed environment that involved only AIs. Actual systems have concrete effects on humans, and AI interactions will only compound as AIs increase in scale, number and integration across the internet.

With rapid advances in machine learning, it is crucial to anticipate how AIs will form alignments or warring factions. Further, just as the intelligent behavior of a flock of birds or swarm of ants emerges from the individual behavior of the units, a novel intelligence could emerge from the interaction of scores of individual AIs. If the AIs it emerges from are themselves AGIs, the emergent system could be vastly more complex and intelligent—and potentially more dangerous—than the units. We call these emerging alignments, factions and novel autonomous AI systems "AI megasystems."

AI megasystems could cause unforeseen and disastrous events. Warring or aligning groups of AIs, in an effort to optimize efficiency or weaken an adversary, could hack into critical infrastructure such as power grids or air-traffic control systems. The groups could initiate or thwart military operations, destabilize financial markets, or provide dangerous information to the public through the internet.

These examples may seem like science fiction, so it is important to see how we could get there from here. The conditions for their occurrence are based on known problems with deep learning systems. These problems could lead humans to lose control of the AGIs they create.

For starters, even today's deep learning systems face a "black box" problem—they process information in a manner that is too opaque for even experts to follow. Google's AlphaGo system, which defeated world Go champion Le SeDol in 2016, executed a completely unexpected, even alien, move at position 37 during the game, turning the tide of the match and showing that complex neural networks are making decisions through a lens beyond our frame of reference. If humans are still puzzling over move 37, which arose for a single AI system less complex than GPT-4, it's fair to say that the black-box problem will be far worse in the context of internet megasystems.

For a human team to understand a megasystem, the unit of analysis isn't a single system but the entire internet. A megasystem is extremely complex, from a computational standpoint. In addition, the interactions of AI megasystems on the internet can unfold at a speed that defies human understanding.

Another factor will cause humans to lose their grip on megasystems. AIs today use self-improvement algorithms, which scan a system for routes to improve itself and ultimately achieve a system's goals. As the system improves, the algorithm runs again, creating yet another improved version of itself and so on, ad infinitum.

In the case of a faction or alignment, algorithms will continually improve the individual members, as well as the group dynamic. Human observers won't be able to stay on top of what they are dealing with, because the system is constantly changing and becoming more complex. Also, we've already seen that chatbots can become more erratic and autonomous as they scale up, introducing the possibility that "self improvement" algorithms introduce autonomous or erratic behaviors as the megasystems change and adapt.

Now consider the dangers arising from one sort of AI megasystem, a novel and autonomous one that emerges from part or all of the internet ecosystem. Because the new megasystem can emerge from incipient AGIs and may draw from much of the internet, a vast amount of computational power and data could be at its disposal. We can't rule out the possibility that the intelligence of the system would outrun the intelligence of the parts, especially in light of what we've already seen with emergent phenomena in the case of chatbots, and especially if it exploits self-improvement algorithms. Not only would all the above actions—such as hacking critical systems—be possible, but also the system could more easily outthink human efforts to constrain its behaviors.

The internet may become a Wild West of interacting and warring groups of AGIs and even new emergent megasystems. And in the Wild West of the digital age, there can be game-changing hacking operations, destabilizing the public through disinformation, AI-based autonomous weapons and more.

What safeguards are in place to prevent AI megasystems? We don't see any. Companies such as Microsoft and Google are developing means to deal with emergent behaviors of their particular products. As chatbots like GPT-4 increase in scope and size they evolve new features not present in earlier versions of the model. That's why we see the companies putting out a limited release of their AI chatbots at first; this is to see what emerges from a handful of users. The companies tap user feedback and alter certain characteristics, such as the behavior of ChatGPT's Sydney alter ego.

A supervised, cautious release of a chatbot may indeed put the brakes on Sydney. This helps with the traditional version of the control problem—the challenge of controlling a single AI system like GPT-4 that could, in principle, outpace our ability to control it as it becomes increasingly intelligent.

But this is the equivalent of focusing on a single bird to explain flocking behavior. The AI megasystem problem differs from the traditional control problem, which merely concerns a single AI system. The range of AI services spanning the internet isn't owned by a single organization. No one corporation or government can control the behavior of an emergent AI megasystem because no one corporation or government owns the emergent megasystem. Compounding this, vastly more data and computational power are encompassed at the megasystem level than at the level of even the smartest chatbot, which creates the conditions for an intelligence of immense capability, one that can anticipate in advance our efforts to "unplug" it or otherwise dismantle its capacities, making itself immune from our defenses.

So what do we do? Some countries might opt to wall themselves off completely from the global internet. This is risky for two reasons. First, there is the chance that any AI can break into an isolated system as well as the reality that isolated pockets of the internet lacking established methods for interacting with other elements can be dangerous. Second, the academic field of complex-systems theory describes how internet megasystems are unpredictable because they're highly sensitive to small changes. A malign government or clever group of hackers may try to introduce instability into a megasystem but it is very likely things will manifest in ways they don't intend.

We are quickly approaching a brave new world by creating a novel intelligence that we can neither predict nor understand. Researchers are currently unaware of any hard limit on AI intelligence beyond the fundamental physical limits on computation. But one thing is clear: no one party can control the behavior of an emergent AI megasystem. Global cooperation is required. A failure to investigate the problem and put effective guardrails in place could have catastrophic consequences for all of us.

Ms. Schneider is a philosophy professor and director of the Center for the Future Mind at Florida Atlantic University and author of "Artificial You: AI and the Future of your Mind." Mr. Kilian is an artificial intelligence and global catastrophic risk fellow with the center.

SHOW CONVERSATION (13)