ARTIFICIAL INTELLIGENCE

Researchers propose test for AI sentience

Human consciousness theories inform preliminary checklist for bots

By **Elizabeth Finkel**

In 2021, Google engineer Blake Lemoine made headlines—and got himself fired—when he claimed that LaMDA, the chatbot he'd been testing, was sentient. Artificial intelligence (AI) systems, especially so-called large language models such as LaMDA and ChatGPT, can certainly seem conscious. But they're trained on vast amounts of text to imitate human responses. So how can we really know?

Now, a group of 19 computer scientists, neuroscientists, and philosophers has come up with an approach: not a single definitive test, but a lengthy checklist of attributes that, together, could suggest but not prove an AI is conscious. In a 120-page discussion paper posted as a preprint this week, the researchers draw on theories of human consciousness to propose 14 criteria, and then apply them to existing AI architectures, including the type of model that powers ChatGPT.

None is likely to be conscious, they conclude. But the work offers a framework for evaluating increasingly humanlike AIs, says co-author Robert Long of the San Francisco–based nonprofit Center for AI Safety. "We're introducing a systematic methodology previously lacking."

Adeel Razi, a computational neuroscientist at Monash University and a fellow at the Canadian Institute for Advanced Research (CIFAR) who was not involved in the new paper, says that is a valuable step. "We're all starting the discussion rather than coming up with answers."

Until recently, machine consciousness was the stuff of science fiction movies such as *Ex Machina*. "When Blake Lemoine was fired from Google after being convinced by LaMDA, that marked a change," Long says. "If AIs can give the impression of consciousness, that makes it an urgent priority for scientists and philosophers to weigh in." Long and philosopher Patrick Butlin of the University of Oxford's Future of Humanity Institute organized two workshops on how to test for sentience in AI.

For one collaborator, computational neuroscientist Megan Peters at the University of California, Irvine, the issue has a moral dimension. "How do we treat an AI based on its probability of consciousness? Personally this is part of what compels me."

Enlisting researchers from diverse disciplines made for "a deep and nuanced exploration," she says. "Long and Butlin have done a beautiful job herding cats."

One of the first tasks for the herd was to define consciousness, "a word full of traps," says another member, machine learning pioneer Yoshua Bengio of the Mila-Quebec Artificial Intelligence Institute. The researchers decided to focus on what New York University philosopher Ned Block has termed "phenomenal consciousness," or the subjective quality of an experience—what it is like to see red or feel pain.

But how does one go about probing the phenomenal consciousness of an algorithm? Unlike a human brain, it offers no signals of its inner workings detectable with an electroencephalogram or MRI. Instead, the researchers took "a theory-heavy approach," explains collaborator Liad Mudrik, a cognitive neuroscientist at Tel Aviv University: They would first mine current theories of human consciousness for the core descriptors of a conscious state, and then look for these in an AI's underlying architecture.

To be included, a theory had to be based on neuroscience and supported by empirical evidence, such as data from brain scans during tests that manipulate consciousness using perceptual tricks. It also had to allow for the possibility that consciousness can arise regardless of whether computations are performed by biological neurons or silicon chips.

Six theories made the grade. One was the Recurrent Processing Theory, which proposes that passing information through feedback loops is key to consciousness. Another, the Global Neuronal Workspace Theory, contends that consciousness arises when independent streams of information pass through a bottleneck to combine in a workspace analogous to a computer clipboard.

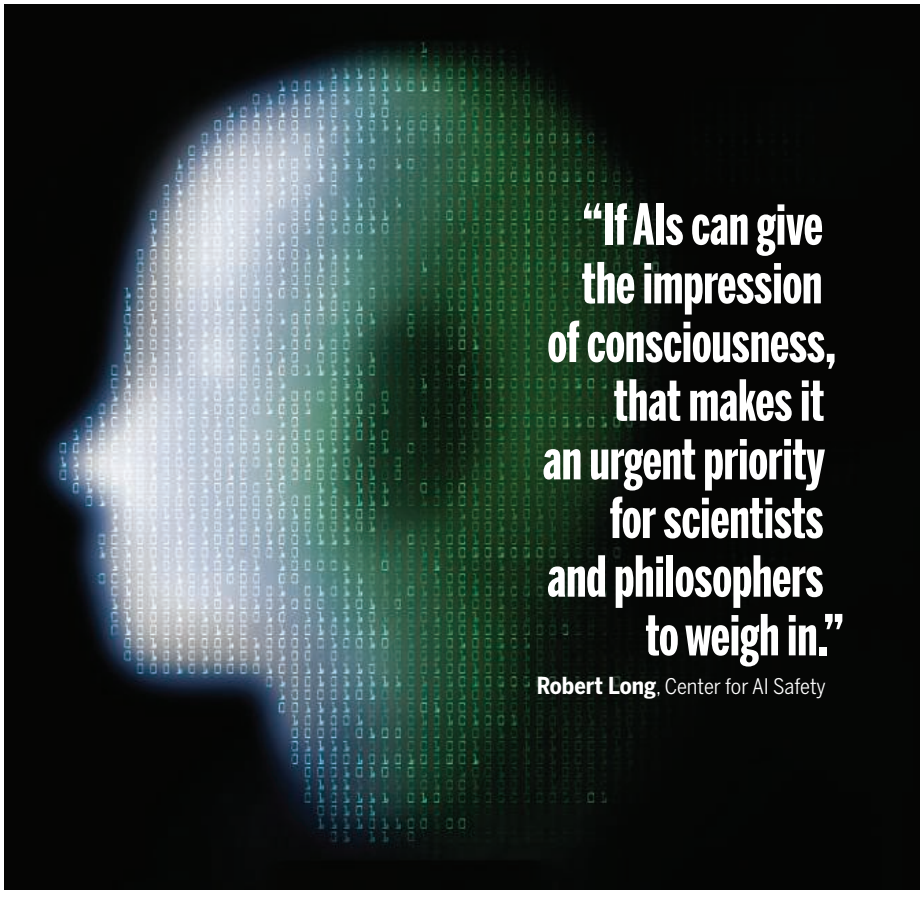

IMAGE: MARK GARLICK/SCIENCE PHOTO LIBRARY VIA GETTY IMAGES

Higher Order Theories suggest consciousness involves a process of representing and annotating basic inputs received from the senses. Other theories emphasize the importance of mechanisms for controlling attention and the need for a body that gets feedback from the outside world. From the six included theories the team extracted their 14 indicators of a conscious state.

The researchers reasoned that the more indicators an AI architecture checks off, the more likely it is to possess consciousness. Mila-based machine learning expert Eric Elmoznino applied the checklist to several AIs with different architectures, including those used for image generation such as Dall-E2. Doing so required making judgment calls and navigating gray areas. Many of the architectures ticked the box for indicators from the Recurrent Processing Theory. One variant of the type of large language model underlying ChatGPT came close to also exhibiting another feature, the presence of a global workspace.

Google's PaLM-E, which receives inputs from various robotic sensors, met the criterion "agency and embodiment." And, "If you squint there's something like a workspace," Elmoznino adds.

DeepMind's transformer-based Adaptive Agent (AdA), which was trained to control an avatar in a simulated 3D space, also qualified for "agency and embodiment," even though it lacks physical sensors like PaLM-E has. Because of its spatial awareness, "AdA was the most likely … to be embodied by our standards," the authors say.

Given that none of the AIs ticked more than a handful of boxes, none is a strong candidate for consciousness, although Elmoznino says, "It would be trivial to design all these features into an AI." The reason no one has done so is "it is not clear they would be useful for tasks."

The authors say their checklist is a work in progress. And it's not the only such effort underway. Some members of the group, along with Razi, are part of a CIFAR-funded project to devise a broader consciousness test that can also be applied to organoids, animals, and newborns. They hope to produce a publication in the next few months.

The problem for all such projects, Razi says, is that current theories are based on our understanding of human consciousness. Yet consciousness may take other forms, even in our fellow mammals. "We really have no idea what it's like to be a bat," he says. "It's a limitation we cannot get rid of." ■

Elizabeth Finkel is a journalist based in Melbourne, Australia.

BIOTECHNOLOGY

Chemists convert electricity into fuel for living cells

ATP generated with renewable power could be used to manufacture proteins and medicines

By **Robert F. Service**

Power plants incessantly burn fossil fuels to convert the solar energy stored by plants eons ago into electricity. But going the other direction—converting electricity into a biologically useful form of energy—has been much more difficult. Now, however, a simple chemical scheme can convert electrical energy into adenosine triphosphate (ATP), the chemical fuel used by all cells, a research team reports. With the process, electricity from renewable sources might someday power biofactories to make everything from protein supplements to medicines.

"This is really exciting," says Michael Jewett, a bioengineer at Stanford University. "This new approach harnesses biological processes to carry out functions that nature never needed, but could benefit society."

Within the cells of plants, organelles called chloroplasts use sunlight to generate ATP as part of the process of photosynthesis. The ATP then powers myriad reactions essential for metabolism. When an ATP molecule is used, it is stripped of one of its phosphate groups, creating adenosine diphosphate, or ADP. The ADP is then recycled and fed more captured energy to regenerate ATP. Plant-eating animals burn glucose to power this same cycle, which occurs some 10 million times per second in every cell.

Industrial biotechnologists tap into this cycle by harnessing specially bred or modified microbes to manufacture everything from biofuels to pharmaceuticals. A process typically starts by growing plants to make sugar or other food that can be fed to yeast, *Escherichia coli*, or other industrial microbes. The microbes use the food to generate ATP that powers the desired biochemical reactions. But plants typically only convert 1% of the energy in sunlight into sugars or other compounds, so such processes are inefficient.

In contrast, solar cells typically convert 20% or more of the energy in sunlight to electricity. Given that disparity, Tobias Erb, a synthetic biologist at the Max Planck Institute for Terrestrial Microbiology, and his colleagues sought a way to convert electricity into ATP more directly. Others had tried before. In 2016, researchers in Spain did so by precisely orienting copies of an ATP-generating enzyme, called ATP synthase, in a membrane adjacent to an electrode. The approach worked in the lab but seemed too complex to be practical, Erb says.

Erb's team set out to devise a simpler approach, creating the "AAA cycle" in which four enzymes in solution harness electricity and use it to convert ADP, added as a reagent, into ATP. Key to the process, Erb says, is a tungsten-containing enzyme called aldehyde ferredoxin oxidoreductase (AOR) that was isolated from a bacterium just 9 years ago. AOR can't directly convert ADP to ATP. Rather it acts a bit like an engine to power the process. "AOR is an energy converter," Erb says.

AOR grabs pairs of electrons from an electrode and uses them to add an energy-rich chemical bond to a starting compound called propionate, converting it to propionaldehyde. Other enzymes then modify that chemical further, until a final enzyme remakes the starting compound, restarting the cycle while unleashing the energy in the bond. "This process releases energy that is used to generate ATP," says Shanshan Luo, a postdoctoral researcher at the Max Planck institute and lead author of the study. The team then used this ATP to drive the conversion of DNA into RNA and proteins in a cell-free setup, a result they reported last week in *Joule*.

"The simple AAA cycle is a clever and elegant approach … that is much simpler than how biology naturally makes ATP," says Drew Endy, a synthetic biologist at Stanford. It could be an enabling technology for the emerging field of electrobiosynthesis, which focuses on using electricity to power the growth of everything from food to pharmaceuticals, he adds. "It's hard for me to overstate the importance of this possibility," Endy says.

The AAA cycle must be improved, however. In solution, AOR survives only for about 1 hour. But Erb says his team is already trying to evolve more stable enzymes and protect them inside a gel they can attach to an electrode. If either succeeds, bioengineers could soon have a new way to power production processes. ■

Researchers propose test for AI sentience

Elizabeth Finkel

Science **381** (6660), . DOI: 10.1126/science.adk4479