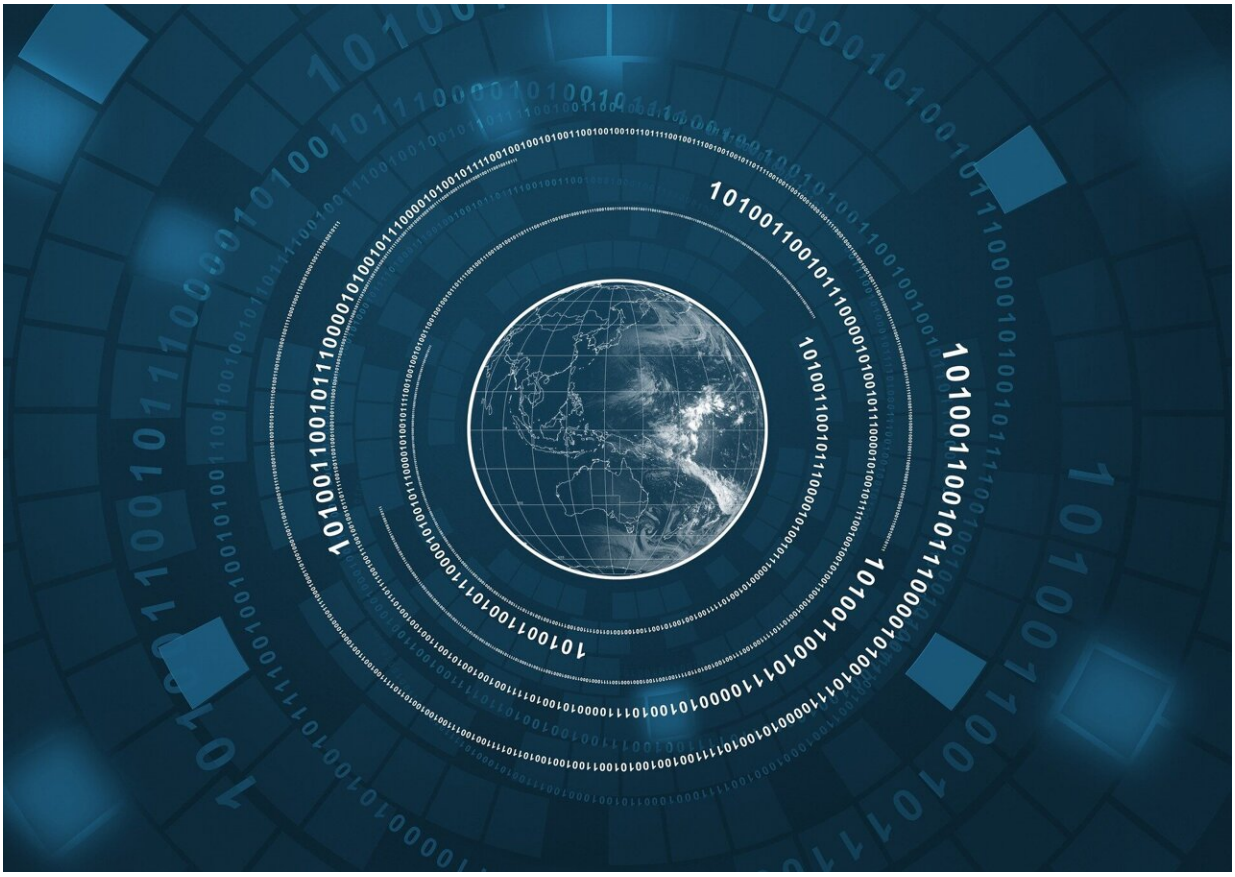
Video: Expert calls for new approach to AI, a 'civilization-ending' technology

April 10 2023, by Rachel Leven



Credit: Pixabay/CC0 Public Domain

AI technology has the power to change the world, said Stuart Russell, a UC Berkeley computer science professor and leading AI expert. It could

improve quality of life for people across the planet or destroy civilization, he said. At an April 5 event, he urged both companies to pivot how they're building AI and countries to regulate AI to ensure it furthers human interests.

"Intelligence really means the power to shape the world in your interests, and if you create systems that are more intelligent than humans either individually or collectively then you're creating entities that are more powerful than us," said Russell at the lecture organized by the CITRIS Research Exchange and Berkeley AI Research Lab. "How do we retain power over entities more powerful than us, forever?"

"If we pursue [our current approach], then we will eventually lose control over the machines. But, we can take a different route that actually leads to AI systems that are beneficial to humans," said Russell. "We could, in fact, have a better civilization."

Releases of chatbots like ChatGPT have provided a lens for the public to learn what AI can do and its future opportunities and dangers. But the chance to change the world—and the $13.5 quadrillion that Russell describes as a "low-ball" estimate for the anticipated value creation from AGI—makes slowing down a high-stakes request.

Existing AI systems like ChatGPT operate in a black box, Russell said. It's unclear whether these tools have goals of their own, if those goals align with ours or whether they can pursue their goals, he said. Instances like the one where a chatbot repeatedly professed its love to a *New York Times* reporter, who rebuffed the bot's advances, suggests they may be able to, he said.

AI should instead be designed to further human interests, to recognize it doesn't know what those interests are, and to seek evidence to identify and act upon those interests, Russell said. This would require a

rethinking of AI concepts like planning, reinforcement learning, and supervised learning that rely on knowing the objective upfront. It also needs to be developed in a "well-founded" way, with a rigorous understanding of all the components and how they work together. That will allow us to predict how these systems will behave, he said.

"I just don't know any other way to achieve enough confidence in the behavior of these systems," Russell said.

Even if technologies are built on this foundation, there must be rules prohibiting the release of unsafe AI, he said.

The last "civilization-ending technology"—atomic energy—has been the subject of intense governance and extreme care on behalf of its engineers, Russell said. Even less charged technology fields, like aviation, are meticulously regulated. AI should be, too, he said.

There's already an international legal framework laying out what responsible AI is and delineating related recommendations, Russell said. Developers should be able to show that AI is robust, predictable and does not present an undue risk to society before it's deployed. Companies should abide by these principles and countries should turn these into rules, he said.

"You should not deploy systems whose internal principles of operation you don't understand, that may or may not have their own internal goals that they are pursuing and that you claim show 'sparks of AGI,'" he said, referring to a recent paper written by Microsoft researchers claiming that OpenAI's GPT-4 shows "sparks of artificial general intelligence."

"If we believe we have sparks of AGI, that's a technology that could completely change the face of the Earth and civilization," said Russell. "How can we not take that seriously?"

Provided by University of California - Berkeley