UC San Diego
UC San Diego News Center

By Anthony King Dec 10, 2020

A Human-Centered Approach to Artificial Intelligence

Institute for Practical Ethics keynote advocates for cross-disciplinary development of groundbreaking technology

The UC San Diego Institute for Practical Ethics presented a new model for artificial intelligence technology Dec. 3, virtually hosting famed AI expert Stuart Russell as their third annual keynote speaker. Russell, former vice-chair of the World Economic Forum's Global Agenda Council on AI and Robotics, advocated for artificial intelligence that takes a human-centered approach, one with the capacity to lift the living standards of everyone on Earth.

Stuart Russell

"How does a machine take actions in the service of our objectives when it doesn't even know what they are?" said Russell, the Smith-Zadeh Chair in Engineering at UC Berkeley and founder of that university's Center for Human-Compatible Artificial Intelligence. "That's a puzzle, but it's a solvable puzzle."

Broadly defined, artificial intelligence is the development of computers to perform tasks that are traditionally completed by humans. Some of the most outrageous examples include cyborg-style robots, but AI already encompasses many aspects of daily life: speech and face recognition, social media algorithms and even password protection on the websites we log into every day.

But the same technological advances that led to these uses may have much bigger implications in the near future, as investment and interest in artificial intelligence grows. Will self-driving cars be safe? Can language barriers be more easily overcome for learning? Should governments expand the use of

autonomous weapons?

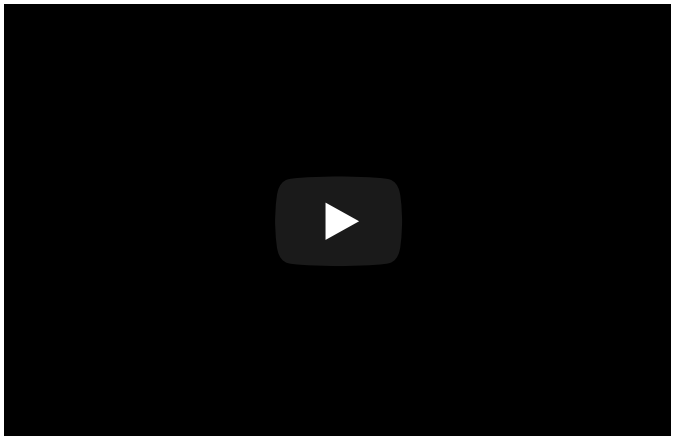
Relying heavily on a humanistic approach

Russell explained his new model, what he calls “provably beneficial AI,” as grounded by three, informal principles: the only objective of the machine is to satisfy human preferences, the machine does not know what those preferences are—an uncertainty he said allows humans to remain in control—and human behavior, through active choice, gives evidence of what those preferences are and will become.

Designers then take these three principles and use them in development, allowing machines to behave very differently than the traditional, standard model of artificial intelligence technology known today: where human preferences do not exist.

Using the example of the self-driving car, a passenger tells the car to take it to the airport and, under the standard model, the car will attempt to achieve this objective at any cost, including not allowing itself to be “turned off” because, Russell explained, this would mean the machine had failed the task.

“In the new model, the thinking goes in a quite different way,” he said: the machine knows it may be turned off if it does something wrong, but it doesn’t know what “wrong” means, and therefore relies on the user to teach it. Optimally, the new model forces machines, robots or algorithms to automatically defer to humans, ask permission before taking action, be “minimally invasive” and empower action in the user by providing more choices.



“With this model, the better the AI, the better the outcome because it’s going to be better able to infer your preferences and better able to satisfy those preferences,” he said.

Russell’s research covers a wide range of topics in artificial intelligence, including machine learning, reasoning, real-time decision making, computational physiology and philosophical foundations. His books include “The Use of Knowledge in Analogy and Induction,” “Do the Right Thing: Studies in Limited Rationality” and “Artificial Intelligence: A Modern Approach.” His latest is “Human Compatible: Artificial Intelligence and the Problem of Control.”

“Professor Stuart Russell is one of the most distinguished and impactful researchers on artificial intelligence in the world, and having our greater community engage with his work provides benefit to everyone involved,” institute co-director Craig Callender said. “What’s especially remarkable about Professor Russell is that he finds time to extend his work outside of academia.”

In addition to the World Economic Forum, Russell has worked with the United Nations on developing a global monitoring system for the nuclear test-ban treaty, and holds fellowships with the American Association for Artificial Intelligence, the Association for Computing Machinery and the American Association for the Advancement of Science.

“We are delighted Dr. Russell was the Institute for Practical Ethics keynote speaker this year,” John H. Evans, institute co-director and associate dean of the Division of Social Sciences, said. “As he pointed out during the public talk, artificial intelligence is yet another important advancement that will benefit from being truly interdisciplinary in its development.”

A practical institute for pressing problems

The Institute for Practical Ethics was established in the Division of Arts and Humanities in late 2017 to address a growing link between advanced technology and real-world application. Co-directors Craig Callender of Philosophy and John H. Evans of Sociology have led the institute as it developed collaborative research in three key areas: data science, genetics and genome technology, and climate-change mitigation.

“From active genetics and genetic engineering, to data science and the environment, the institute has developed into a ‘living hub’ of researchers, visiting and faculty experts, post-doc scholars and Ph.D. fellows committed to questions about the ethics and social implications of some of the most groundbreaking science of our lives,” Arts and Humanities Dean Cristina Della Coletta said.

The mission of the institute is to ensure the impacts of research will include and benefit all people, and was founded in part because UC San Diego’s first leaders also cared deeply about the implications of groundbreaking science, including Roger Revelle and Herbert York.

UC San Diego Executive Vice Chancellor for Academic Affairs Elizabeth H. Simmons opened the keynote address, acknowledging the university as being at the cutting edge of discoveries that impact a global society.

“Our global ecosystem of innovation is growing ever larger and more complex, and, paradoxically, both more interdependent and yet more fragmented than ever before,” she said. “In particular, as we continue to advance our technological capabilities, we must, at the same time, consider the ethical implications of such innovations.”

Keep up with campus news by subscribing to *This Week @ UC San Diego*