HUMAN-COMPATIBLE ARTIFICIAL INTELLIGENCE

STUART RUSSELL (UC BERKELEY)

FRIDAY, MARCH 11TH • 2-4 PM • BAKER 180-101

Abstract: I will briefly survey recent and expected developments in AI and their implications. Some are enormously positive, while others, such as the development of autonomous weapons and the replacement of humans in economic roles, may be negative. Beyond these, one must expect that AI capabilities will eventually exceed those of humans across a range of real-world-decision making scenarios. Should this be a cause for concern, as Alan Turing, Elon Musk, Stephen Hawking, and others have suggested? And, if so, what can we do about it? While some in the mainstream AI community dismiss the issue, I will argue that the problem is real and that the technical aspects of it are solvable if we replace current definitions of AI with a version based on provable benefit to humans. This, in turn, raises a host of questions with which the social sciences and humanities have wrestled for centuries.

Bio: Stuart Russell is a Professor of Computer Science at the University of California at Berkeley, holder of the Smith-Zadeh Chair in Engineering, and Director of the Center for Human-Compatible AI. He is a recipient of the IJCAI Computers and Thought Award and held the Chaire Blaise Pascal in Paris. In 2021 he received the OBE from Her Majesty Queen Elizabeth and gave the Reith Lectures. He is an Honorary Fellow of Wadham College, Oxford, an Andrew Carnegie Fellow, and a Fellow of the American Association for Artificial Intelligence, the Association for Computing Machinery, and the American Association for the Advancement of Science. His book "Artificial Intelligence: A Modern Approach" (with Peter Norvig) is the standard text in AI, used in 1500 universities in 135 countries. His research covers a wide range of topics in artificial intelligence, with a current emphasis on the long-term future of artificial intelligence and its relation to humanity. He has developed a new global seismic monitoring system for the nuclear-test-ban treaty and is currently working to ban lethal autonomous weapons.

technology, policy & ethics
LECTURE SERIES AT CAL POLY, SLO

WITH SUPPORT FROM
CLA, PHIL, & CENG

For more information, contact Dr. Ava Thomas Wright, avwright@calpoly.edu