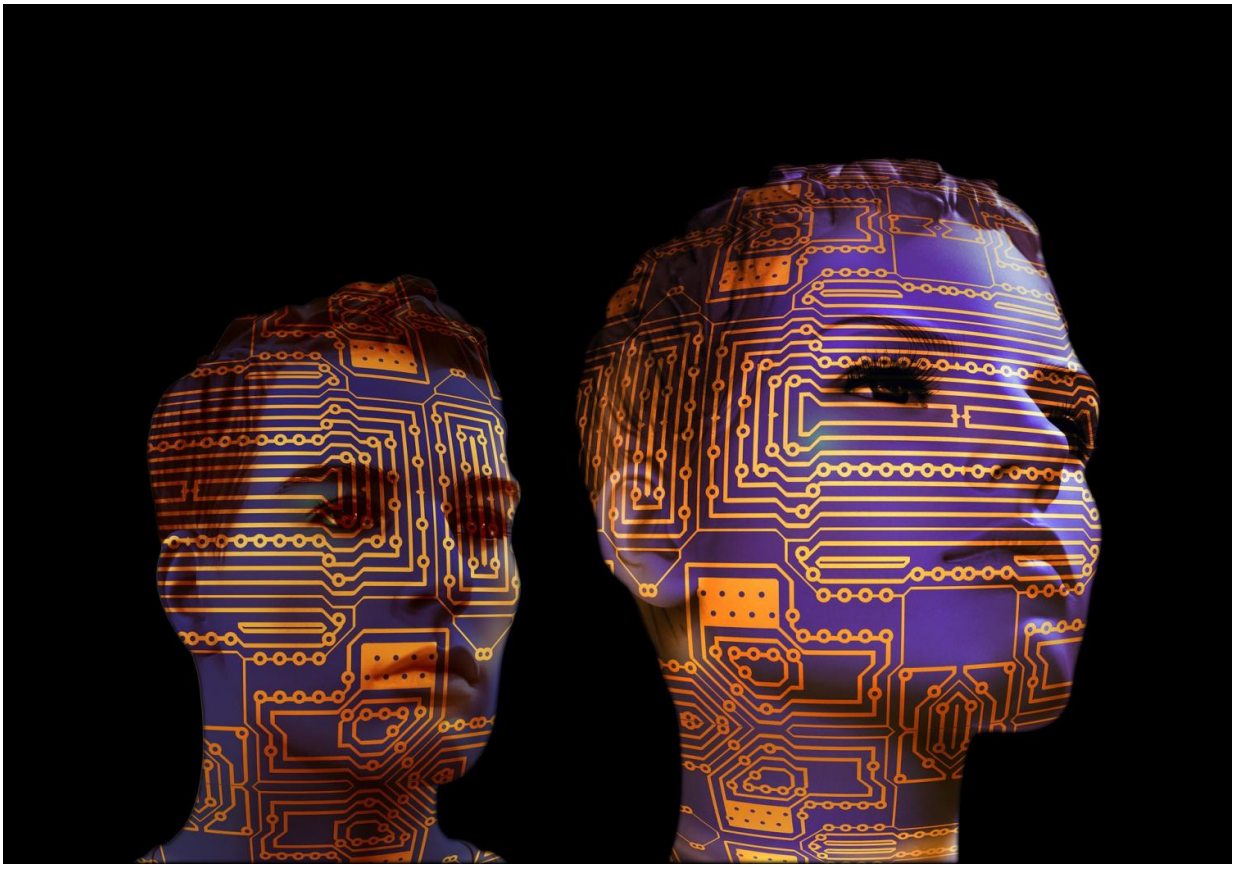
Teaching human values to artificial intelligences

September 8 2016, by Bill Steele



Credit: CC0 Public Domain

Two Cornell experts in artificial intelligence (AI) have joined a nationwide team setting out to ensure that when computers are running the world, they will make decisions compatible with human values.

"We are in a period in history when we start using these machines to make judgments," said Bart Selman, professor of computer science. "If decisions are properly structured, the horrors we've seen in the movies won't happen."

Selman and Joseph Halpern, professor of computer science, have become co-principal investigators for the Center for Human-Compatible Artificial Intelligence, a nationwide research effort based at the University of California, Berkeley. Initially they will collaborate with scientists at Berkeley and the University of Michigan. Soon the team expects to add experts in economics, philosophy and social sciences.

The primary focus of the new center is to ensure that AI systems are beneficial to humans, said Stuart Russell, a Berkeley professor of electrical engineering and computer science. The center will work on ways to guarantee that the AI systems of the future, which may be entrusted with control of critical infrastructure and may provide essential services to billions of people, will act in a manner that is aligned with human values.

"Systems are already being fielded in society," Selman said. "We must make sure the robotic systems actually know about human ethics and human values."

Much of Selman's research is in the area of computer science called "decision theory." He recently worked on a project funded by Tesla Motors CEO Elon Musk to make self-driving cars safer, and that includes problems in decision-making. The car must decide if it's worth the risk to pass the slow-moving car up ahead. Ultimately this could evolve into the moral dilemmas debated by philosophers, like the "Trolley Problem." A runaway trolley will crash and kill five people, but you can stop it by pushing a man off a bridge so he lands on the tracks, probably killing him in the process. As a less theoretical problem, should

your self-driving car be protective of you at the expense of other drivers?

The stakes will be even higher when computers manage air traffic control or the power grid, or make medical decisions in hospitals.

Halpern also works with decision theory. Typically, when making a decision, there is uncertainty about what will happen, he points out. How many people may die in the trolley crash? How seriously will the guy on the tracks be injured?

One solution to uncertainty is to have more data, he noted, and part of the answer lies in giving computers access to Big Data. "If you have lots of data you can estimate the probabilities well and get a much better handle on uncertainty," he explained. Some of that may be data on human behavior: Russell has suggested robots should learn about human decision-making by observing human activity.

"As we go into the world with massive AI and robots," Halpern concluded, "how should we prepare ourselves?"

Provided by Cornell University