Empowering robots for ethical behavior

July 18 2017

Credit: CC0 Public Domain

Scientists at the University of Hertfordshire in the UK have developed a concept called Empowerment to help robots to protect and serve humans, while keeping themselves safe.

Robots are becoming more common in our homes and workplaces and this looks set to continue. Many robots will have to interact with humans

in unpredictable situations. For example, self-driving cars need to keep their occupants safe, while protecting the car from damage. Robots caring for the elderly will need to adapt to complex situations and respond to their owners' needs.

Recently, thinkers such as Stephen Hawking have warned about the potential dangers of artificial intelligence, and this has sparked public discussion. "Public opinion seems to swing between enthusiasm for progress and downplaying any risks, to outright fear," says Daniel Polani, a scientist involved in the research, which was recently published in *Frontiers in Robotics and AI*.

However, the concept of "intelligent" machines running amok and turning on their human creators is not new. In 1942, science fiction writer Isaac Asimov proposed his three laws of robotics, which govern how robots should interact with humans. Put simply, these laws state that a robot should not harm a human, or allow a human to be harmed. The laws also aim to ensure that robots obey orders from humans, and protect their own existence, as long as this doesn't cause harm to a human.

The laws are well-intentioned, but they are open to misinterpretation, especially as robots don't understand nuanced and ambiguous human language. In fact, Asimov's stories are full of examples where robots misinterpreted the spirit of the laws, with tragic consequences.

One problem is that the concept of "harm" is complex, context-specific and is difficult to explain clearly to a robot. If a robot doesn't understand "harm", how can they avoid causing it? "We realized that we could use different perspectives to create 'good' robot behavior, broadly in keeping with Asimov's laws," says Christoph Salge, another scientist involved in the study.

The concept the team developed is called Empowerment. Rather than

trying to make a machine understand complex ethical questions, it is based on robots always seeking to keep their options open. "Empowerment means being in a state where you have the greatest potential influence on the world you can perceive," explains Salge. "So, for a simple robot, this might be getting safely back to its power station, and not getting stuck, which would limit its options for movement. For a more futuristic, human-like robot this would not just include movement, but could incorporate a variety of parameters, resulting in more human-like drives."

The team mathematically coded the Empowerment concept, so that it can be adopted by a robot. While the researchers originally developed the Empowerment concept in 2005, in a recent key development, they expanded the concept so that the robot also seeks to maintain a human's Empowerment. "We wanted the robot to see the world through the eyes of the human with which it interacts," explains Polani. "Keeping the human safe consists of the robot acting to increase the human's own Empowerment."

"In a dangerous situation, the robot would try to keep the human alive and free from injury," says Salge. "We don't want to be oppressively protected by robots to minimize any chance of harm, we want to live in a world where robots maintain our Empowerment."

This altruistic Empowerment concept could power robots that adhere to the spirit of Asimov's three laws, from self-driving cars, to robot butlers. "Ultimately, I think that Empowerment might form an important part of the overall ethical behaviour of robots," says Salge.

More information: Christoph Salge et al, Empowerment As Replacement for the Three Laws of Robotics, *Frontiers in Robotics and AI* (2017). DOI: 10.3389/frobt.2017.00025

Provided by Frontiers