Comment on John Searle's "What Your Computer Can't Know"

Bill Hibbard

20 September 2014

This is a brief comment on John Searle's article "What Your Computer Can't Know" in the 9 October 2014 issue of the New York Review of Books. His article is a review of two books: "The 4th Revolution: How the Infosphere Is Reshaping Human Reality" by Luciano Floridi, and "Superintelligence: Paths, Dangers, Strategies" by Nick Bostrom, both published by the Oxford University Press.

Regarding Bostrom's book Searle writes:

> This is why the prospect of superintelligent computers rising up and killing us, all by themselves, is not a real danger. Such entities have, literally speaking, no intelligence, no motivation, no autonomy, and no agency.

He also writes:

> Why is it so important that the system be capable of consciousness? Why isn't appropriate behavior enough? Of course for many purposes it is enough. If the computer can fly airplanes, drive cars, and win at chess, who cares if it is totally nonconscious? But if we are worried about a maliciously motivated superintelligence destroying us, then it is important that the malicious motivation should be real. Without consciousness, there is no possibility of its being real.

Later in the article, regarding of "the project of creating an artificial brain that does what real human brains do," Searle writes:

> To carry out such a project it is essential to remember that what matters are the inner mental processes, not the external behavior. If you get the processes right, the behavior will be an expression of those processes, and if you don't get the processes right, the behavior that results is irrelevant.

I believe Searle has this backwards. The danger of artificial intelligence is in its behavior, and whether it is conscious or possesses other attributes of human thought is irrelevant. Computers that can drive cars and fly airplanes certainly pose dangers to humans and in fact Google has gone to great efforts to design safety into their self-driving cars. Computers that can run the entire world economy and provide constant companionship to all humans will pose great danger to humans.

Malicious motivation is irrelevant to many of the dangers posed by super-intelligent machines. There are two forms of "wireheading" to guard against: computers that delude themselves about their observations of the environment and computers that modify the source of approval for their actions, for example modifying humans. There are also dangers from what Omohundro described as "Basic AI Drives." Super-intelligent machines may be tools of

competition among humans, who will be careless about these dangers because they are caught up in the heat of competition.