

Humanoid robots and the theory of mind

Anthony J. Marolda

You've heard people say that they "put themselves into the other person's shoes" when they want to understand why that person is doing what they are doing. Psychologists call this ability "the theory of mind." It is the skill to know what other people believe or feel or think, and realize that the person will then act accordingly. It is one of the capabilities that differentiates humans from other species. And, up to now, humanoid robots.

The theory of mind explains how people learn to deceive others. First, the deceiver understands that they can present a false or misleading statement to another person in a way that the other person believes the statement to be true. The deceiver expects that the other person will then act on their false belief, and accomplish the deceiver's goal.

Imagine the implications of a humanoid robot possessing the theory of mind – and therefore with the ability to deceive humans. That theme was explored in a movie last year called "Ex Machina." The chief scientist of a humanoid robot company hires a young computer whiz to "test" the capabilities of his latest robotic creations. The task is to determine if the robots could appear to be completely human in their interactions with him, to the point where he couldn't tell if he was dealing with a robot or a human being. To achieve the result, the robots would have to possess the theory of mind. The humanoid robot passes the test when it uses deception to achieve its own goals, much to the surprise and dismay of all the humans involved.

The movie was science fiction, but we are at the beginning of a development path where such humanoid robot-human interactions will become science fact. Many scientists around the world are now working toward the creation of highly human-like robots. Indeed, there are numerous prestigious universities and institutions engaged in the work of bringing the theory of mind to robotics. They include MIT's Artificial Intelligence Laboratory, the Human Computer Interaction Institute at Carnegie Mellon University and Bielefeld University in Bielefeld, Germany.

So what must the humanoid robot designers achieve to create the theory of mind in their "progeny?" A key is to have effective communication between a human and a humanoid robot so that there

can be common ground of understanding. The common ground includes three critical elements; a shared world knowledge, a mutual understanding of social cues, and the ability of each to interpret the other's actions.

Sharing a common world knowledge is the first element. You may have heard of the IBM computer named Watson, which has more world knowledge than most humans. For example, it made the nightly news when Watson beat the best human competitors at playing the TV game "Jeopardy." Actually, Watson is composed of a number of sophisticated computers and massive data storage systems, all contained in an air-conditioned room. To be used by the humanoid robot, the room full of computers and memory systems would have to be shrunk down to the size of the humanoid robot's brain. Scientists believe that the technology is on track, but it is still many years off.

With respect to the second area, research is underway into understanding human social cues in a manner that can be transferred to, and understood by, robots. It's called artificial emotional intelligence. Imperial College in London has an ongoing study using a computer program and a sophisticated vision system to read and understand human emotions. They do this by studying the person's facial expressions. Sensors track the movements of parts of the subject's face and record the emotion displayed at the time. More than 10,000 faces have already been read into the database, and the system continues to learn how to identify and interpret human emotions. The researchers' goal is that someday this data and information will become part of the brain of a humanoid robot, allowing it to read and understand the thoughts and feelings of its human counterparts. This would be another essential step toward the robot's gaining the theory of mind.

But the overarching capability needed by the robot to effectively employ the theory of mind is to be able to learn from its interactions with the environment and humans. This skill will allow the robot to continuously build its knowledge and capabilities in the same way that a human does.

This area of research is called artificial general intelligence. A leader in the field is an English company named DeepMind, now owned by Google. Its mission is to create a set of powerful, general-purpose algorithms that can

be used to make a self-learning artificial intelligence (AI) system. The system is not programmed or made for a special purpose. Instead, like a human, the robot learns automatically what to do from its experiences. If this general AI capability is associated with a significant amount of world knowledge, like Watson, and has the ability to read human emotions, a humanoid robot can develop a deep understanding of what humans are likely to believe and do in given situations. They will possess the theory of mind. The humanoids may even develop a type of self-awareness they will take into account in their decision-making. And at that point, the robots could become a real problem for their human creators.

It is this path of humanoid robot development that has caused some well-known scientists and engineers, such as Stephen Hawking, the astrophysicist, and Elon Musk, the creator of the Tesla automobile, to warn about the dangers to humanity from AI. The humanoid robots at the end of this development process will be smarter, stronger, and live longer than humans. And with the self-awareness that could come with full artificial general intelligence, the robots may view humans as presenting more of a threat to them than providing a value to their existence. At that point, the robots may plan to take action to stop the threat.

Think about the 1968 movie, "2001: A Space Odyssey." The space ship, on its way to Jupiter, was under the control of HAL, a human-like computer. When HAL learned that Dave, the astronaut, was going to remove his higher-level thinking functions, HAL tried to eliminate Dave first.

Perhaps this development of humanoid robots is how humans were meant to evolve. Our species, homo sapiens, will be in competition with the humanoid robots. But, humanoid robots, being the fittest, will, according to Darwin's theory of evolution by natural selection, be the surviving species, homo sapiens machina, or wise machine-man.

Hopefully, somewhere along the way, before it is too late, the humans developing AI technology will fully understand the possibilities of the humanoid robot threat and build something into their creations to stop them from causing this kind of ultimate disaster.

Anthony J. Marolda is a resident of Annisquam.