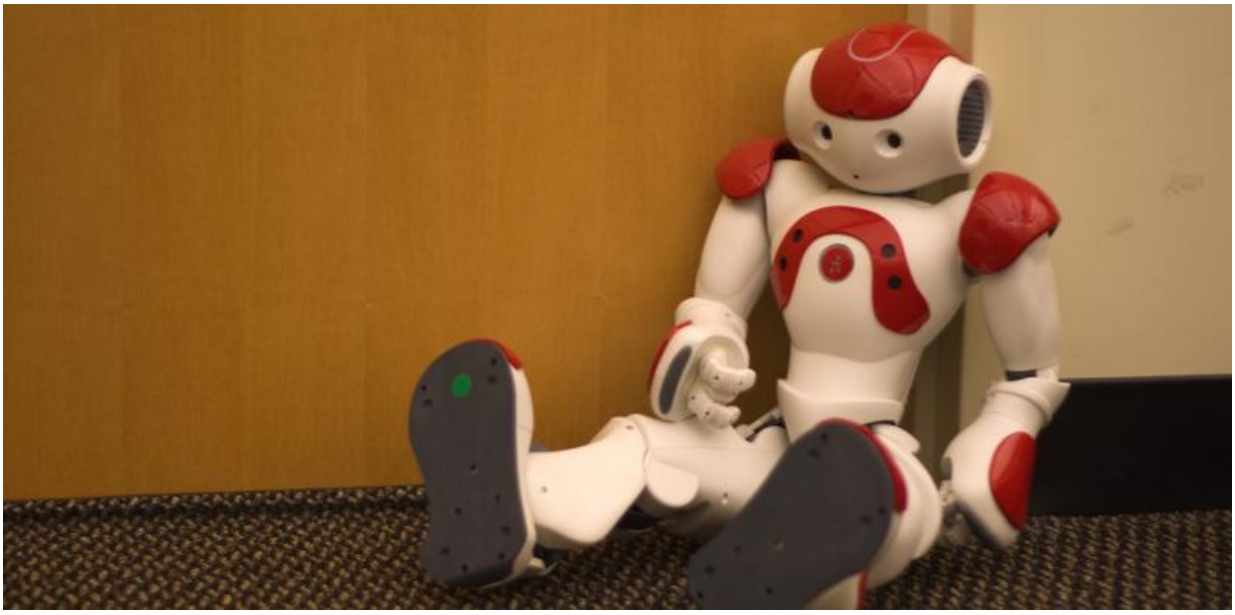
How to design trustworthy robot butlers that we won't want to treat like humans

May 13 2016, by Rob Wortham



Nao - a robot created for companionship. Credit: Jiuguang Wang/wikimedia, CC BY-SA

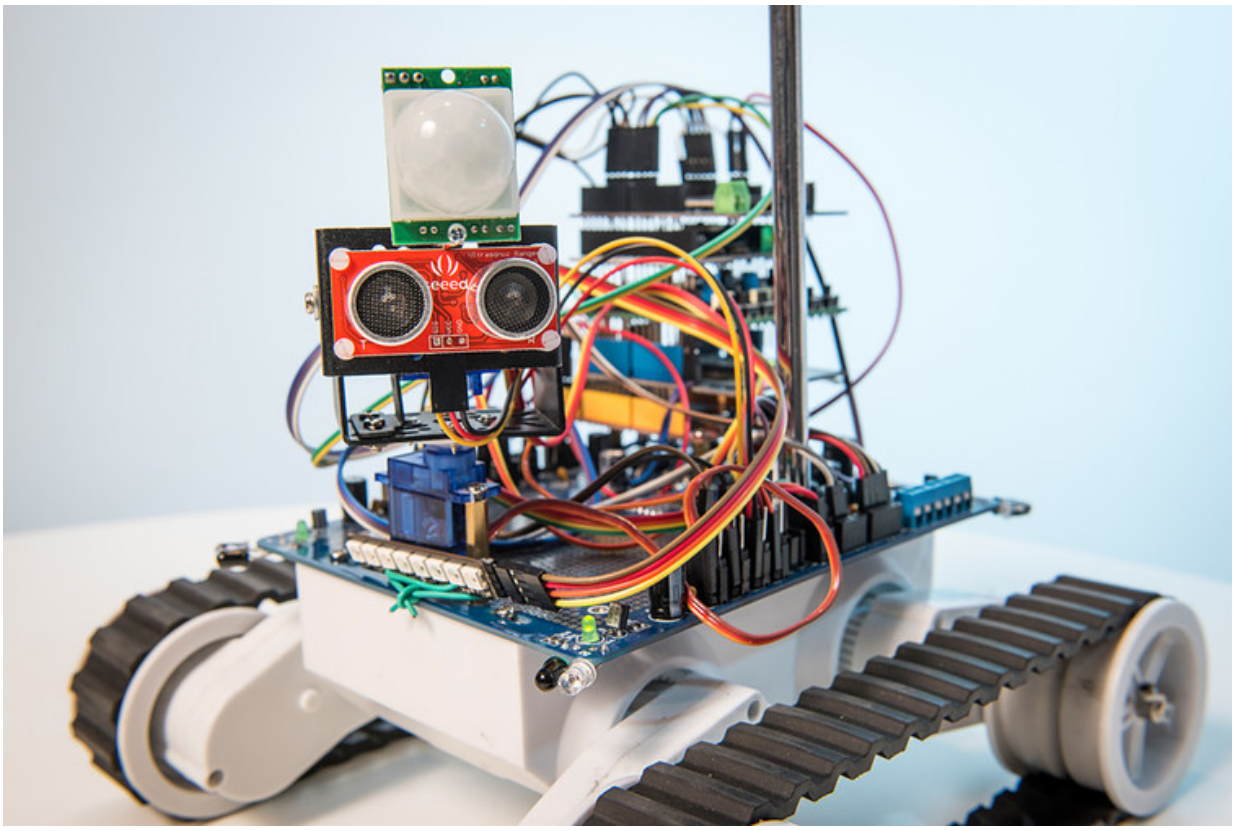
Does your car "not want" to start on cold mornings? And does your toaster "like" burning your toast? This kind of intentional language is natural to us and built into the way we interact with the world – even with machines. This is because we have evolved to become extremely social animals, understanding others by forming mental models of what they are thinking. We use these skills to understand the behaviour of

anything complicated we interact with, especially robots.

It's almost like we believe that machines have minds of their own. And the fact that we perceive them as intelligent is partly why they have such potential. Robots are moving beyond industrial, commercial and scientific applications, and are already used in hospitals and care homes. Soon it will be normal to interact with robots in our daily lives, to complete useful tasks. Robots are also being used as companions, particularly for elderly patients with cognitive impairment such as dementia. After years of scientific study, this has proven very successful at improving long-term quality of life.

However, there are ethical concerns about vulnerable people forming relationships with machines, in some cases even believing them to be animals or people. Are robot designers intending to deceive patients? As robots become more important to us, how can we trust them not to mislead us, indeed should we trust them at all?

In 2010, a group of academics produced ethical guidelines for how we should build robots, much like science fiction writer Isaac Asimov's famous laws. Asimov stated that robots could not do anything to harm a human being; that a robot should always obey a human; and that a robot should defend itself so long as this didn't interfere with the first two rules. Similarly, these academics produced the EPSRC Principles of Robotics. For me, the most interesting principle is number four: "Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent."

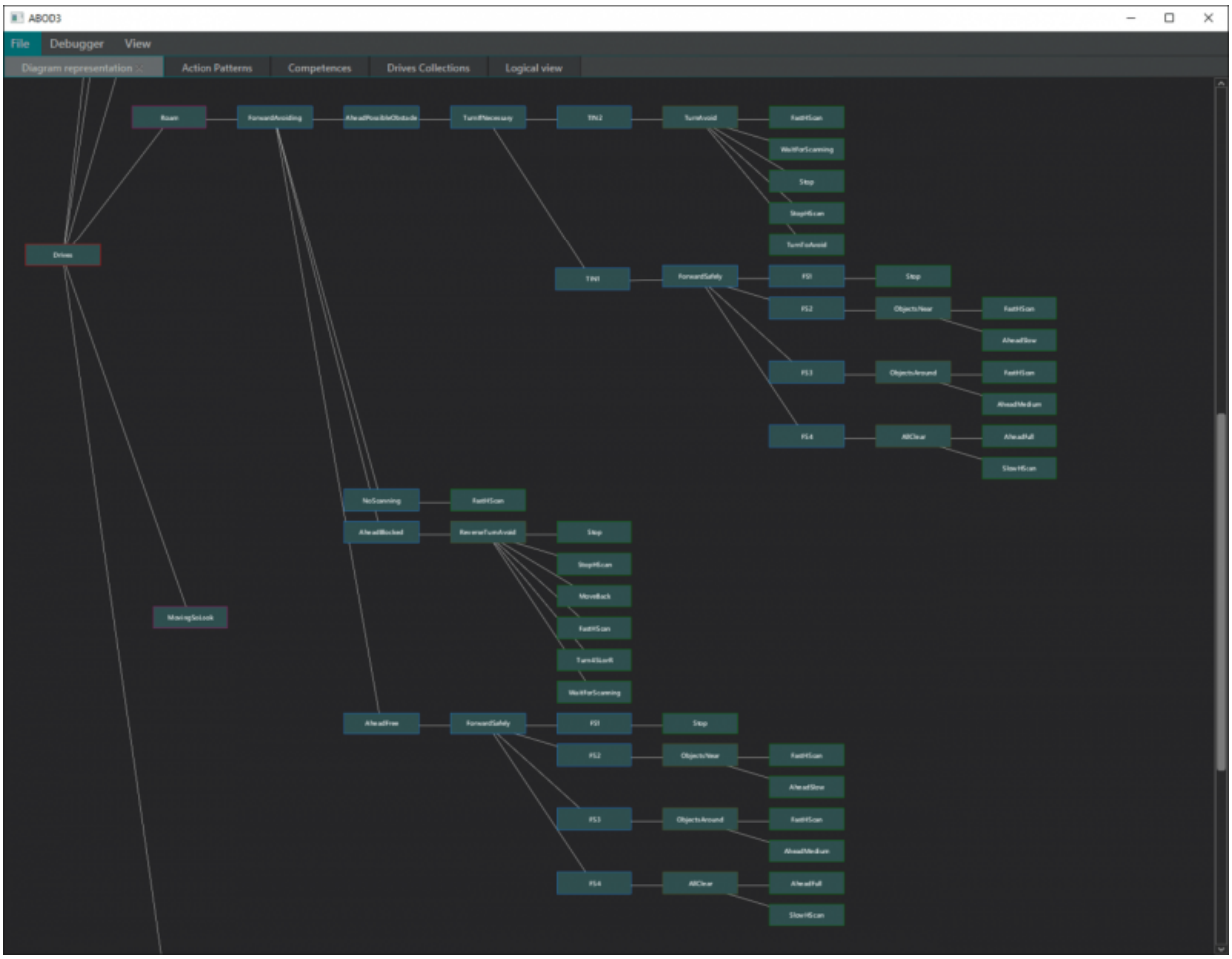


Meet the human-tracking robot. Credit: University of Bath., Author provided

Hidden decisions

In my research group, we are conducting experiments with robots to investigate how well we understand them. Ultimately, we want to find out how we can best design robots to improve our mental models. We're trying to show that it's possible to create transparent robots that are useful, and when necessary emotionally engaging, despite having a transparent machine nature. We assert that if we can make robots more transparent, then we won't need to trust them, we'll always know what they are doing.

We use a simple robot that moves around a room avoiding objects while searching for humans. When it finds a human it flashes lights, does a small wiggle dance, and then trundles off seeking another one. Sometimes when in a corner it goes to sleep to save battery. That's all it does. We videoed the robot, showed this to a group of 22 people and asked them what the robot was doing and why.



Screen shot of the display showed to the second group. Author provided

Some of the answers were remarkable. Based on cues from the

environment, and the imaginations of the people, they came up with all sorts of ideas about what the robot was up to – views that were generally quite wrong. For instance there is a bucket in the room, and several people were sure the robot was trying to throw something into it. Others noticed an abstract picture in the room and wondered if the robot was going to complete the picture. These people were mainly graduates in professional jobs, and several had science, technology, engineering or maths degrees. Almost all used computers every day. Although we did not program the robot, nor create the room explicitly to mislead, the observers were deceived.

We showed the same video to an almost identical second group. However this group was simultaneously shown a display demonstrating each decision in the robot's action selection system controlling its behaviour, synchronised with the robot moving around the room and tracking objects. It is a kind of dynamic heat map of the processes and decisions being made inside the robot's brain, making plain the focus of attention of the robot, and outlining the steps it takes to achieve its goals.

The results, to be published at the International Joint Conference on Artificial Intelligence in July, were remarkable. The second group came to a much better understanding of what the robot was doing, and why. We expected that result. What we didn't expect was that the second group were twice as sure that the robot was "thinking". It seems that an improved mental model of the robot is associated with an increased perception of a thinking machine, even when there is no significant change in the level of perceived intelligence. The relationship between the perception of intelligence and thinking is therefore not straightforward.

This is encouraging as it shows that we can have robots that are transparently machines and yet are still engaging, in that the participants attribute them with intelligence and thinking. We are beginning to show

that designers can create something appealing without the need to hide the true capabilities of the robot.

So, the robot butler of the future may have transparency built in. Perhaps you can ask it what its doing and it will tell you by showing you or talking about what's going on in its brain. We'd like to see that mechanism built in to the low level brain code of the robot, so it has to tell it like it is. It would be nice for the user to be able to dial this up or down depending on how familiar they are with the tasks the robot is doing.

We plan to try other ways of making robots transparent using speech technology, and combinations of graphics, text and speech, and we hope to be able to produce more detailed guidelines and better tools for robot designers to help them build robots we don't need to trust.

This article was originally published on The Conversation. *Read the* original article.

Source: The Conversation