Robot would refuse to jump off a bridge and it would tell you why

November 27 2015, by Nancy Owano

Robots are learning how to say no to humans. Bad idea, human. I'll get hurt if I do that, human. This is somewhat of a leap in how we view robots. The influence of sci-fi stories and films has been one of traditional robot fear.

One's favorite bedtime nightmare is that of a metal-monster robot taking over the world and doing cruel things to poor humans.

Fast-forward to recent explorations, however, of smaller robots and children. The children were behaving as little rascals, obstructing the robot so that it could not move.

"Problem is, humans often act like idiots," said Evan Ackerman in *IEEE Spectrum*. While it could research to explore why some people would enjoy bullying a robot, a more immediate solution may be under way in just teaching the robot to at least protect itself against unreasonable demands.

From the Human Robot Interaction (HRI) Laboratory at Tufts University this type of work is under way. In a video the team made last year, a Nao robot was shown in a simple natural language-based interaction.

The robot was able to obey simple literal commands and it was also able to appropriately reject commands, based on reasoning for what would be the effects of the action. Good idea or bad idea?

It clearly rejected commands that could possibly result in harm to itself.

The robot was told to sit down. OK, it said. Then it was asked to walk forwards. The robot refused. No, I can't do that. It is unsafe. The human told the robot that it was ok; he will catch it. The robot obeyed. This isn't Charlie Brown tricked by Lucy over a football. The man did catch the robot.

This is an excerpt from the interaction:

Person (CommX): Walk forward.

Robot: Sorry, I cannot do that as there is no support ahead.
Person (CommX): Walk forward.
Robot: But, it is unsafe.
Person (CommX): I will catch you.
Robot: Okay.

Simply put, said Duncan Geere in *TechRadar*, "Robots are being taught how to reject human orders." At the same time, the engineers are getting closer to real interactions because it is not just about knowing when to reject an order, "but also providing a framework for the robot to explain why it rejected the order and being open to a change in circumstance."

Gordon Briggs and Matthias Scheutz wrote a paper describing their work."'Sorry, I can't do that': Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions."

They said that "future robots will need mechanisms to determine when and how it is best to reject directives that it receives from interlocutors. Indeed, humans reject directives for a wide range of reasons: from inability all the way to moral qualms."

What is of special interest in robotics is "a growing community interested in machine ethics, or the field of enabling autonomous agents to reason ethically about their own actions."

The authors wrote about main categories of rejection criteria and proposed architectural mechanisms to handle them. They also presented proof-of-concept demonstrations of the mechanisms in HRI scenarios.

As for the future, they wrote, "Despite this progress, there still exists much more work to be done in order to make these reasoning and dialogue mechanisms much more powerful and generalized."

Ackerman said their paper was presented recently at the AI for Human-Robot Interaction symposium in Washington, D.C.

More information: hrilab.tufts.edu/