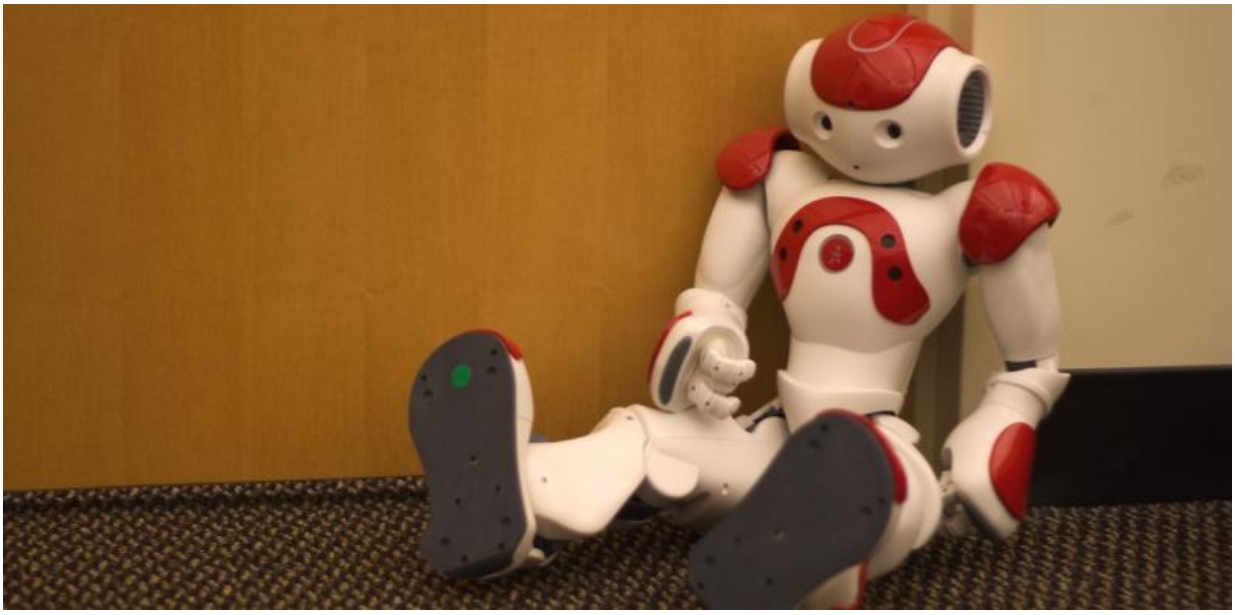
Why robots need to be able to say 'No'

April 8 2016, by Matthias Scheutz



Is this robot refusing a human order? Credit: Jiuguang Wang, CC BY-SA

Should you always do what other people tell you to do? Clearly not. Everyone knows that. So should future robots always obey our commands? At first glance, you might think they should, simply because they are machines and that's what they are designed to do. But then think of all the times you would not mindlessly carry out others' instructions – and put robots into those situations.

Just consider:

An elder-care robot tasked by a forgetful owner to wash the "dirty clothes," even though the clothes had just come out of the washerA preschooler who orders the daycare robot to throw a ball out the windowA student commanding her robot tutor to do all the homework instead doing it herselfA household robot instructed by its busy and distracted owner to run the garbage disposal even though spoons and knives are stuck in it.

There are plenty of benign cases where robots receive commands that ideally should not be carried out because they lead to unwanted outcomes. But not all cases will be that innocuous, even if their commands initially appear to be.

Consider a robot car instructed to back up while the dog is sleeping in the driveway behind it, or a kitchen aid robot instructed to lift a knife and walk forward when positioned behind a human chef. The commands are simple, but the outcomes are significantly worse.

How can we humans avoid such harmful results of robot obedience? If driving around the dog were not possible, the car would have to refuse to drive at all. And similarly, if avoiding stabbing the chef were not possible, the robot would have to either stop walking forward or not pick up the knife in the first place.

In either case, it is essential for both autonomous machines to detect the potential harm their actions could cause and to react to it by either attempting to avoid it, or if harm cannot be avoided, by refusing to carry out the human instruction. How do we teach robots when it's OK to say no?

How can robots know what will happen next?

In our lab, we have started to develop robotic controls that make simple

inferences based on human commands. These will determine whether the robot should carry them out as instructed or reject them because they violate an ethical principle the robot is programmed to obey.

Telling robots how and when – and why – to disobey is far easier said than done. Figuring out what harm or problems might result from an action is not simply a matter of looking at direct outcomes. A ball thrown out a window could end up in the yard, with no harm done. But the ball could end up on a busy street, never to be seen again, or even causing a driver to swerve and crash. Context makes all the difference.

It is difficult for today's robots to determine when it is okay to throw a ball – such as to a child playing catch – and when it's not – such as out the window or in the garbage. Even harder is if the child is trying to trick the robot, pretending to play a ball game but then ducking, letting the ball disappear through the open window.

Explaining morality and law to robots

Understanding those dangers involves a significant amount of background knowledge (including the prospect that playing ball in front of an open window could send the ball through the window). It requires the robot not only to consider action outcomes by themselves, but also to contemplate the intentions of the humans giving the instructions.

To handle these complications of human instructions – benevolent or not – robots need to be able to explicitly reason through consequences of actions and compare outcomes to established social and moral principles that prescribe what is and is not desirable or legal. As seen above, our robot has a general rule that says, "If you are instructed to perform an action and it is possible that performing the action could cause harm, then you are allowed to not perform it." Making the relationship between obligations and permissions explicit allows the robot to reason through

the possible consequences of an instruction and whether they are acceptable.

In general, robots should never perform illegal actions, nor should they perform legal actions that are not desirable. Hence, they will need representations of laws, moral norms and even etiquette in order to be able to determine whether the outcomes of an instructed action, or even the action itself, might be in violation of those principles.

While our programs are still a long way from what we will need to allow robots to handle the examples above, our current system already proves an essential point: robots must be able to disobey in order to obey.

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

Source: The Conversation

Citation: Why robots need to be able to say 'No' (2016, April 8) retrieved 25 April 2024 from https://phys.org/news/2016-04-robots.html