Researchers suggest adding self-doubt to robots to keep them from overstepping bounds

June 7 2017, by Bob Yirka

Credit: CC0 Public Domain

(Tech Xplore)—A team of researchers from the University of California at Berkeley has found some evidence that suggests making robots less self-assured might make them easier to integrate into society. In their

paper uploaded to the prepress server *arXiv*, the group explains their theory and the results of a simulation they ran to test it.

As artificial intelligence and robot technology improve, we humans are going to be faced with some decisions—one, the authors suggest, could be how much self-confidence to give them. Too much, and they might try to override our wishes; too little might make them less than useful.

To better understand the problem, the researchers created a mathematical model meant to depict an interaction between a robot and a human with parameters for adjusting self-confidence. In one simulation, a robot with a built-in off switch was asked to perform a desired task. At that point, a human was given the option of allowing the robot to continue or hit its off switch. But the robot had the ability to override its own off switch, and thus the human's wishes. The researchers found that robots given a healthy dose of self-confidence, as might be expected, chose to override the human's wishes and turn themselves back on. When given just a little bit of self-confidence, on the other hand, the robot stayed off, even if it judged itself doing a good job.

In the real world, the researchers offer an example of what we humans already face—the newsfeed on Facebook. What started out as a means of offering news that a bot thought would be useful has become a major nuisance, because it instead offers a constant stream of fake news. This could have been averted, the researchers contend, if the bot had less confidence.

On the other hand, the researchers note, we have to be careful to not take away too much confidence—a robot might need to override a child's wishes, for example, when they are challenged to stop or change routes. One way to give robots the right amount of self-confidence, they suggest, might be to give them more information to work with. One

example would be a robot that makes coffee only in the morning, because it has learned that is when coffee is wanted.

More information: The Off-Switch Game, arXiv:1611.08219 [cs.AI] arxiv.org/abs/1611.08219

Abstract

It is clear that one of the primary tools we can use to mitigate the potential risk from a misbehaving AI system is the ability to turn the system off. As the capabilities of AI systems improve, it is important to ensure that such systems do not adopt subgoals that prevent a human from switching them off. This is a challenge because many formulations of rational agents create strong incentives for self-preservation. This is not caused by a built-in instinct, but because a rational agent will maximize expected utility and cannot achieve whatever objective it has been given if it is dead. Our goal is to study the incentives an agent has to allow itself to be switched off. We analyze a simple game between a human H and a robot R, where H can press R's off switch but R can disable the off switch. A traditional agent takes its reward function for granted: we show that such agents have an incentive to disable the off switch, except in the special case where H is perfectly rational. Our key insight is that for R to want to preserve its off switch, it needs to be uncertain about the utility associated with the outcome, and to treat H's actions as important observations about that utility. (R also has no incentive to switch itself off in this setting.) We conclude that giving machines an appropriate level of uncertainty about their objectives leads to safer designs, and we argue that this setting is a useful generalization of the classical AI paradigm of rational agents.