Asimov's Laws won't stop robots harming humans, so we've developed a better solution

July 11 2017, by Christoph Salge

Credit: Andrea Piacquadio from Pexels

How do you stop a robot from hurting people? Many existing robots, such as those assembling cars in factories, shut down immediately when a human comes near. But this quick fix wouldn't work for something like a self-driving car that might have to move to avoid a collision, or a care robot that might need to catch an old person if they fall. With robots set to become our servants, companions and co-workers, we need to deal with the increasingly complex situations this will create and the ethical and safety questions this will raise.

Science fiction already envisioned this problem and has suggested various potential solutions. The most famous was author Isaac Asimov's Three Laws of Robotics, which are designed to prevent robots harming humans. But since 2005, my colleagues and I at the University of Hertfordshire, have been working on an idea that could be an alternative.

Instead of laws to restrict robot behaviour, we think robots should be empowered to maximise the possible ways they can act so they can pick the best solution for any given scenario. As we describe in a new paper in Frontiers, this principle could form the basis of a new set of universal guidelines for robots to keep humans as safe as possible.

The Three Laws

Asimov's Three Laws are as follows:

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

While these laws sound plausible, numerous arguments have demonstrated why they are inadequate. Asimov's own stories are arguably a deconstruction of the laws, showing how they repeatedly fail in different situations. Most attempts to draft new guidelines follow a similar principle to create safe, compliant and robust robots.

Credit: AI-generated image (disclaimer)

One problem with any explicitly formulated robot guidelines is the need to translate them into a format that robots can work with. Understanding the full range of human language and the experience it represents is a very hard job for a robot. Broad behavioural goals, such as preventing harm to humans or protecting a robot's existence, can mean different things in different contexts. Sticking to the rules might end up leaving a robot helpless to act as its creators might hope.

Our alternative concept, empowerment, stands for the opposite of helplessness. Being empowered means having the ability to affect a situation and being aware that you can. We have been developing ways to translate this social concept into a quantifiable and operational technical language. This would endow robots with the drive to keep their options open and act in a way that increases their influence on the world.

When we tried simulating how robots would use the empowerment principle in various scenarios, we found they would often act in surprisingly "natural" ways. It typically only requires them to model how the real world works but doesn't need any specialised artificial intelligence programming designed to deal with the particular scenario.

But to keep people safe, the robots need to try to maintain or improve human empowerment as well as their own. This essentially means being protective and supportive. Opening a locked door for someone would increase their empowerment. Restraining them would result in a short-term loss of empowerment. And significantly hurting them could remove their empowerment altogether. At the same time, the robot has to try to maintain its own empowerment, for example by ensuring it has enough power to operate and it does not get stuck or damaged.

Robots could adapt to new situations

Using this general principle rather than predefined rules of behaviour would allow the robot to take account of the context and evaluate scenarios no one has previously envisaged. For example, instead of always following the rule "don't push humans", a robot would generally avoid pushing them but still be able to push them out of the way of a falling object. The human might still be harmed but less so than if the robot didn't push them.

In the film I, Robot, based on several Asimov stories, robots create an oppressive state that is supposed to minimise the overall harm to humans by keeping them confined and "protected." But our principle would avoid such a scenario because it would mean a loss of human empowerment.

While empowerment provides a new way of thinking about safe robot behaviour, we still have much work to do on scaling up its efficiency so it can easily be deployed on any robot and translate to good and safe behaviour in all respects. This poses a very difficult challenge. But we firmly believe empowerment can lead us towards a practical solution to the ongoing and highly debated problem of how to rein in robots' behaviour, and how to keep robots -– in the most naive sense -– "ethical."

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

Provided by The Conversation

Citation: Asimov's Laws won't stop robots harming humans, so we've developed a better solution (2017, July 11) retrieved 3 May 2024 from https://techxplore.com/news/2017-07-asimov-laws-wont-robots-humans.html