Why Asimov's three laws of robotics are unethical

Aaron Sloman
School of Computer Science, University of Birmingham
http://www.cs.bham.ac.uk/~axs/

This document is available online here, in two formats:
https://www.cs.bham.ac.uk/research/projects/cogaff/misc/asimov-three-laws.html
https://www.cs.bham.ac.uk/research/projects/cogaff/misc/asimov-three-laws.pdf
Originally written in 2006 (possibly earlier!) and updated a few times.
Last updated: 2 Mar 2022
(Extended section at end on related sites.)

Every now and again I get asked for views on Asimov's laws of robotics. Here are some questions and my answers.

Should we be afraid of what intelligent machines might do to us?

Whenever journalists etc. ask me about that my answer is something like this:

> It is very unlikely that intelligent machines could possibly produce more dreadful behaviour towards humans than humans already produce towards one other -- both at individual levels and at social or national levels -- all over this planet, even in the supposedly most civilised and advanced countries,
>
> Moreover, the more intelligent the machines are the less likely they are to produce all the dreadful human behaviours motivated by religious intolerance, nationalism, racialism, greed, and sadistic enjoyment of the suffering of others.
>
> They will have far better goals to pursue.

What do you think about the three laws?

The three laws are

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

  Asimov later added a 'zeroth' law
- 0: A robot may not injure humanity, or, through inaction, allow humanity to come to harm.

For more on the three laws see The Wikipedia entry.

I have always thought these are pretty silly: they just express a form of racialism or speciesism.

If the robot is as intelligent as you or I, has been around as long as you or I, has as many friends and dependents as you or I (whether humans, robots, intelligent aliens from another planet, or whatever), then there is no reason at all why it should be subject to any ethical laws that are different from what should constrain you or me.

My 1978 book, The Computer Revolution in Philosophy had an epilogue that mentioned the need for a society for the liberation of robots.

Current robots and other machines are not very intelligent. They may perform better than humans at narrowly defined tasks, but they don't 'scale-out' as humans do, linking different kinds of knowledge and expertise creatively in solving problems.

So in the majority of cases we should just think about robots as we think about other products of human engineering. The responsibilities of the designers of robots are no different from the responsibilities of other designers of complex systems that are potentially harmful, e.g. nuclear power plants, dangerous chemical plants, giant dams that disrupt ecosystems, weapons that can be put to evil uses, or which can accidentally cause disasters, etc.

In comparison with those, the majority of threats from robots in the foreseeable future will be tiny, partly because AI progress is so hard, for reasons I have been writing about for some time, e.g.
http://www.cs.bham.ac.uk/research/cogaff/
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/AREADME.html
http://www.cs.bham.ac.uk/research/projects/cosy/papers/
http://www.cs.bham.ac.uk/research/cogaff/talks/

Note added 3 Mar 2022
In the 44 years since I wrote *The Computer Revolution in Philosophy* (1978) there has been a vast amount of progress in AI and Robotics. But computer-based intelligent systems still lack kinds of spatial intelligence found in squirrels, young children, ancient mathematicians who discovered constructions, theorems and proofs in geometry, elephants, and many other intelligent animals.

Are the laws implementable ?

There is a prior question as to whether the laws *should* be implemented in robot designs and various social practices, as indicated above. I would regard that as unethical in some cases.

The main obstacle to implementation is vagueness in the laws. E.g. what counts as 'harm'. There are many things that are regarded as good parental behaviour towards their children that might be regarded as harm if done by a stranger (e.g. fiercely telling off, or forcibly restraining, a child who has started to do something potentially very dangerous to him/herself).

Another obstacle involves potential contradictions as the old utilitarian philosophers found centuries ago: what harms one may benefit another, etc., and preventing harm to one individual can cause harm to another. There are also conflicts between short term and long term harm and benefit for the same individual. There is nothing in Asimov's formulation about how the contradictions should be resolved, though I know he and others have noticed the problems and explored some options. (You can look up the history of 'utilitarianism' as an ethical theory if you want to know more. Major early exponents were Jeremy Bentham and John Stuart Mill, but many have challenged or tried to extend their theories, often showing up great difficulties.)

But my main answer remains: humans are among the things on earth to be most feared. If we can find ways to educate and socialise them so that we no longer produce human monsters, or monster cultures, then the same methods should apply to machines with human-like intelligence.

Another law for robots and humans

I recommend the following, which I learnt from the writings of Karl Popper.

> *Always be prepared to consider the possibility that you've got something wrong and there's a better answer than anything you've produced or encountered so far.*

That principle should, of course, be applied to itself.

Related sites

- Lee McCauley on 'The 3 Laws of Robotics'
- Hykel Hosni interviews Julia Staffel in *The Reasoner, Volume 16, Number 2 March - April 2022*, pp 9-14. Asimov's "Laws of robotics" are discussed briefly on page 13. http://blogs.kent.ac.uk/thereasoner/files/2022/03/TheReasoner-162.pdf

The above is not intended to be an exhaustive list: use search engines to find more!

Maintained by Aaron Sloman
School of Computer Science
The University of Birmingham