The hermeneutical intractability of Asimovâ??s three laws of robotics

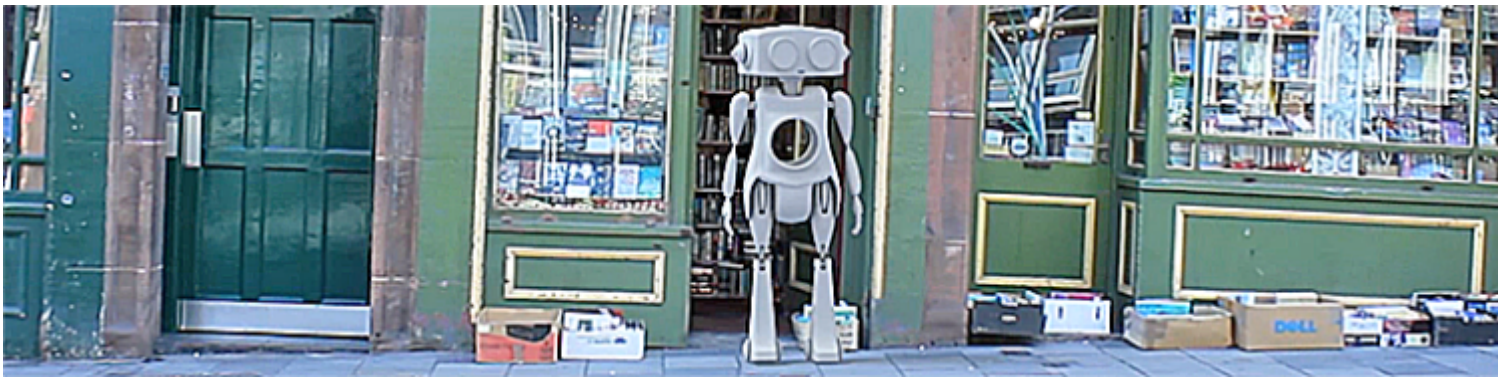
Description

In his sci-fi detective mystery, *I, Robot*, Isaac Asimov writes: â??Powellâ??s radio voice was tense in Donovanâ??s ear: â??Now, look, letâ??s start with the three fundamental Rules of Robotics â?? the three rules that are built most deeply into a robotâ??s positronic brain.â?? The rules follow.

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given 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 Laws.

These rules are coded into the neural fabric of every robot of the future. It means you will always trust your robots to do the right thing. Of course, the story posits an exceptional robot that breaks the rules, and is suspected of murdering a human being.

Asimovâ??s protagonist explains: â??The conflict between the various rules is ironed out by the different positronic potentials in the brain. Weâ??ll say that a robot is walking into danger and knows it. The automatic potential that Rule 3 sets up turns him back. But suppose you order him to walk into that danger. In that case, Rule 2 sets up a counterpotential higher than the previous one and the robot follows orders at the risk of existence.â?•



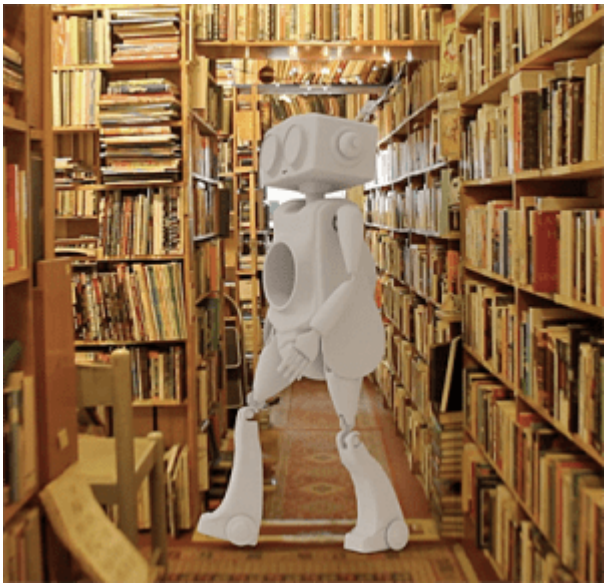
Robot agents

We donâ??t currently ascribe to machines and software the idea of responsibility, culpability, and guilt. In order to shift our moral compass towards Asimovâ??s rule-bound sci-fi scenario we would have to stop attributing responsibility to the inventors, designers, programmers, developers and controllers of our machines, and re-direct it to the machines themselves.

According such autonomy (agency) to machines is not beyond the realms of possibility argues Ugo Pagallo in a book on the legal implications of robotics. One day â??robots would be held responsible for the damages caused by themâ?• â?? and the robot, not the people who put it in place. The issue comes to a head as we think of military robotic drones that are not just remote controlled, but are programmed, or instructed, to seek out a strategic target. He proposes that â??the more robotics advances and becomes more sophisticated, the more likely it is that such machines will need a legal regime of their own.â?•

Re-ordering the rules

In various sci-fi scenarios the sentiment in rule 3 shifts priority to the same level as rule 1, and that sparks the very human contest between self preservation (rule 3) and the prohibition against harm to others (rule 1). Such a shift seems to define the moment of sentience, human-like intelligence, emotional awareness, and consciousness for HAL in *2001: A Space Odyssey* (1968), the runaway replicants in *Blade Runner* (1982), the rogue robot in the film version of *I, Robot* (2004), the maverick synths in the television series *Humans* (2015), and countless other robot inspired narratives. Itâ??s the conflict between the 3 rules that makes us human. Contrariwise, the separation of human ethical behaviour into slavish adherence to such rules would turn us humans into automata â?? or slaves â?? if thatâ??s ever likely. (The â??Unsullied,â?• the elite army of harshly trained eunuchs in *Game of Thrones Season Three* (2013) comes to mind.)



Applying the rules

Shifting the order of the rules brings to light the problem of how such rules get applied in concrete situations. But the rules pose problems in their application irrespective of their ordering. Pagallo highlights the issue of application, and the difficulty in the case of Asimovâ??s rules: â??their abstract and general nature, entails the difficult task of applying Asimovâ??s laws in a given context.â?• But I think that however precise and detailed the wording of any rules (even if coded in formal machine logic) the devil is in the application, as it is for any rule in any legal framework. How far into the future would a robot have to see, and how much information would it need, to be sure that a human being was embarking on a path that would cause her harm â?? and then intervene? Most robot stories of course play with the intractability of such scenarios, and that creates the drama.

Lest we think that autonomous robots have a place in the world beyond fiction, Hans-Georg Gadamer makes the case for the primacy of *application* over slavish implementation of rules: â??Application is neither a subsequent nor a merely occasional part of the phenomenon of understanding, but codetermines it as a whole from the beginningâ?• (285). â??Understandingâ?• is a more cogent human capability to deal with than â??intelligence.â?• I donâ??t think thereâ??s as yet a research program based on *artificial understanding* (AU). But that is what would be required to realise humanoid robots of the kind envisage in science fiction. In so far as human behaviour is influenced at all by rules, itâ??s in their application that understanding arises. To understand, which is to say, to be human, is to be in the world, in community, at work or play, engaged and interacting, and there are no rules for that.

References

- Gadamer, Hans-Georg. 1975. *Truth and Method*. Trans. Joel Weinsheimer. New York: Seabury Press. Originally published in German in 1960.
- Pagallo, Ugo. 2013. *The Laws of Robots: Crimes, Contracts, and Torts*. Dordrecht, Germany: Springer.
- Snodgrass, Adrian, and Richard Coyne. 2006. *Interpretation in Architecture: Design as a Way of Thinking*. London: Routledge.

Notes

- I left out the page numbers for some of the quotes as I was accessing online, unpaginated versions of the texts.
- Robot images are by Lin Youqing, student in the MSc in Design and Digital Media 2014-15.
- See article: Artificial Intelligence Ethics a New Focus at Cambridge University (3 Dec 2015)

Category

1. Nature

Tags

1. AI
2. ethics
3. hermeneutics
4. robotics

Date Created
November 21, 2015

Author
rcoyne99