The Three Laws of Robotics

Machine ethics (or machine morality) is a part of the ethics of AI concerned with the moral behavior of artificially intelligent beings. Machine ethics contrasts with roboethics, which is concerned with the moral behavior of *humans* as they design, construct, use and treat such beings. Machine ethics should not be confused with computer ethics, which focuses on professional behavior towards computers and information.

Isaac Asimov considered the issue in the 1950s in his novel: *I, Robot*. He proposed the Three Laws of Robotics to govern artificially intelligent systems. Much of his work was then spent testing the boundaries of his three laws to see where they would break down, or where they would create paradoxical or unanticipated behavior. His work suggests that no set of fixed laws can sufficiently anticipate all possible circumstances.

Asimov's laws are still mentioned as a template for guiding our development of robots ; but given how much robotics has changed, since they appeared, and will continue to grow in the future, we need to ask how these rules could be updated for a 21st century version of artificial intelligence.

Asimov's suggested laws were devised to protect humans from interactions with robots. They are:

- 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

As mentioned, one of the obvious issues is that robots today appear to be far more varied than those in Asimov's stories, including some that are far more simple. So we need to consider whether we should have a threshold of complexity below which the rules might not be required. It is difficult to conceive a robotic vacuum cleaner having the capability of harming humans or even requiring an ability to obey orders. It is a robot with a single task that can be predetermined prior to it being switched on.

At the other end of the spectrum, however, are the robots designed for military combat environments. These devices are being designed for spying, bomb disposal or load-carrying purposes. These would still appear to align with Asimov's laws, particularly as they are being created to reduce risk to human lives within highly dangerous environments. But it is only a small step to assume that the ultimate military goal would be to create armed robots that could be deployed on the battlefield. In this situation, the First Law – not harming humans – becomes hugely problematic. The role of the military is often to save the lives of soldiers and civilians but often by harming its enemies on the battlefield. So the laws might need to be considered from different perspectives or interpretations.

There's also the question of what counts as harming a human being. This could be an issue when considering the development of robot babies in Japan, for example. If a human were to adopt one of these robots it might arguably cause emotional or psychological harm. But this harm may not have come about from the direct actions of the robot or become apparent until many years after the human-robot interaction has ended.

Intelligence.org
TheConversation.com
Wikipedia.com
[edited]