Erik Persson
Department of Philosophy, Lund University, Sweden
erik.persson@fil.lu.se

Maria Hedlund
Department of Political Science, Lund University, Sweden
maria.hedlund@svet.lu.se

The Trolley problem and Isaac Asimov's First Law of Robotics

The two most commonly stated phrases in connection with AI and ethics are probably "The Three Laws of Robotics" and "The trolley problem". The two are often discussed independently of each other but the trolley problem is also regularly brought forward as an unsolvable problem for Asimov's three laws, and especially for the first law.

The First law of robotics as stated by Isaac Asimov first appeared in the short story *Liar* in 1941 and was formally stated as the first of three in the short story *Runaround* from 1942.

The First Law tells us that "A robot may not injure a human being or, through inaction, allow a human being to come to harm."

The trolley problem is a scenario where you as a by-stander witness an out of control trolley approaching a group of people who are unable to get off the track. If hit by the trolley they will all be killed. Your only option as a by-stander is to either do nothing or shift the trolley onto another track where the trolley will instead hit and kill one person.

The deontological solution to the trolley problem is to not interfere. Doing so would be to actively kill someone, which is unacceptable. The consequentialist solution is the opposite, that is, to save the many even if it means actively killing the few.

The First Law has a deontological character in that it tells us that there is one thing a robot is never allowed to do, namely to injure a human being. On the other hand, it also has a teleological character by banning a certain outcome, that is that human beings come to harm. This double character is an obvious problem when dealing with situations such as the trolley problem that are constructed with the explicit purpose of highlighting the differences between deontological and consequentialist moral theories and intuitions. The trolley problem thus appears unsolvable by the First Law, since a robot is not allowed to choose either solution. It is not allowed to intervene since it would mean injuring a human being but it is also mandatory for the robot to intervene since it is not allowed to let a human come to harm through inaction. What is a poor robot to do?

In this paper, we discuss the following solutions suggested in the literature and their practical and theoretical consequences:

1. Assume that the set of moral dilemmas illustrated by the trolley problem is quite small and inability to handle these cases should not be an objection against using The Three Laws in other situations.

2. Refer to Asimov's suggested "Zeroeth Law of Robotics" stating that: "A robot may not harm humanity or, through inaction, allow humanity to come to harm". Asimov himself suggests that this law implies that the good of the many outweighs the good of the few.

3. Allow for a human operator to strengthen or weaken the different principles depending on context.

4. Construct an algorithm that balances the relative strengths of the deontological and teleological parts of the law based on certain criteria (e.g. the relative number of people on each track or whether some of the involved humans have put the others in danger, etc.).