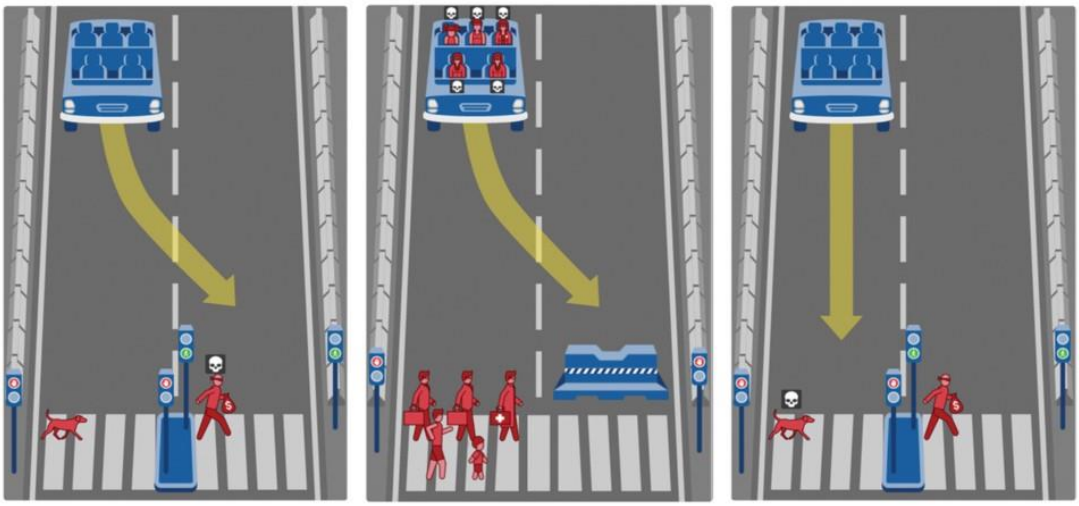
Artificial Intelligence and Ethics

As robots become more autonomous, the notion of computer-controlled machines facing ethical decisions is moving out of the realm of science fiction and into the real world. Society needs to find ways to ensure that they are better equipped to make moral judgments.

The ethics of artificial intelligence is the part of the ethics of technology specific to robots and other artificially intelligent beings. It is typically divided into ***roboethics*** and ***machine ethics***. The term roboethics refers to the morality of how humans design, construct, use and treat robots and other artificially intelligent beings. It considers both how artificially intelligent beings may be used to harm humans and how they may be used to benefit them.

What is more crucial, at the moment, is that even the narrow AI applications that exist today require our urgent attention in the ways in which they are making moral decisions in practical day-to-day situations. For example, this is relevant when algorithms make decisions about who gets access to loans or when self-driving cars have to calculate the value of a human life in hazardous traffic situations.



Teaching morality to machines is hard because humans can't objectively convey morality in measurable metrics that make it easy for a computer to process. In fact, it is even questionable whether we, as humans have a sound understanding of morality at all that we can all agree on. In moral dilemmas, humans tend to rely on gut feeling instead of elaborate cost-benefit calculations. Machines, on the other hand, need explicit and objective metrics that can be clearly measured and optimized.

But in real-life situations, optimization problems are vastly more complex. For example, how do you teach a machine to algorithmically maximize fairness or to overcome racial and gender biases in its training data? A machine cannot be taught what is fair unless the engineers designing the AI system have a precise conception of what fairness is.

Joseph Weizenbaum argued in 1976 that AI technology should not be used to replace people in positions that require respect and care, such as any of these: a soldier, a judge, a nursemaid …etc Weizenbaum explains that we require authentic feelings of empathy from people in these positions. If machines replace them, we will find ourselves alienated, devalued and frustrated. Artificial intelligence, if used in this way, represents a threat to human dignity. Weizenbaum argues that the fact that we are entertaining the possibility of machines in these positions suggests that we have experienced an "atrophy of the human spirit that comes from thinking of ourselves as computers."

Machines cannot be assumed to be inherently capable of behaving morally. Humans must teach them what morality is, how it can be measured and optimized. For AI engineers, this may seem like a daunting task. After all, defining moral values is a challenge mankind has struggled with throughout its history. If we can't agree on what makes a moral human, how can we design moral robots? Nevertheless, the state of AI research and its applications in society require us to finally define morality and to quantify it in explicit terms. This is a difficult but not impossible task. Engineers cannot build a "Good Samaritan AI", as long as they lack a formula for the Good Samaritan human.

Economist.comOecd-forum.org
Weforumorg
[Edited]