Innocence over utilitarianism: Heightened moral standards for robots in rescue dilemmas

April 11 2023, by Mari Peltonen

Credit: AI-generated image (disclaimer)

The Moralities of Intelligent Machines research group headed by Michael Laakasuo investigates people's moral views on imaginary rescue situations where the rescuer is either a human or a robot specifically designed for the task. The rescuer has to decide whether to save, for

example, one innocent victim of a boating accident or two individuals whose irresponsible behavior caused the accident.

"It's about putting the number of lives saved at odds with giving priority to the innocent," says the study's main author Jukka Sundvall.

The goal of the study was to collect data on the factors that people emphasize in their moral assessments about difficult decision-making situations, and whether the emphasis changes if the decision is assigned to a robot. In other words, are robot rescuers expected to adhere to different priorities than humans?

The most important finding in the study was that study participants (N = 3,752) and the respondents of an online survey (in Finnish) conducted with the Finnish public broadcaster Yle (N = 19,000) emphasized the innocence of those who were to be saved in the accident situation more than their number. In general, people thought it was better to save one innocent person than two who had caused the accident, be the rescue agent human or robot. Respondents wanted the rescuer to maximize the number of lives saved only in situations where all parties involved in the accident were equally culpable or not culpable.

Another finding was that this emphasis was highlighted in the case of robots: if the rescuer decided to maximize the number of lives saved by rescuing those responsible for the accident, the decision was more strongly condemned in the case of robot than human rescuers.

Robots are assessed more critically than humans

"Based on the findings, it appears that robots' decisions are assessed on stricter moral criteria," Michael Laakasuo says.

"While robots and humans are subjected to similar moral expectations,

robots are expected to be better than humans at meeting those expectations."

One possible reason for this is that people wish for automated decision-making to be "right" much more often than people. If this does not happen, it calls into question the purpose of the automation.

"Perhaps poor moral decisions made by humans can be seen as understandable incidents, while in the case of robots they are considered indicators of errors in programming," Sundvall muses.

Examining attitudes towards new technology

On the practical level, stricter moral criteria can result in markedly negative reactions in real-life situations where the outcome of automated decision-making is morally poor from citizens' perspectives. High expectations may hinder the deployment of automated decision-making. It is not always clear in advance what the general public considers to be a morally worse option in individual circumstances, let alone the quantity of morally poorer outcomes considered an "acceptable number of mistakes."

The study is part of the fields of moral psychology and human–technology interaction studies, and its purpose is to expand our understanding of moral thinking and attitudes towards new technologies.

According to Sundvall, the study is important because the development of artificial intelligence and robotics is currently topical.

"The possibilities for automated decision-making in various sectors of society are increasing, and it's useful to try to anticipate related problems," Sundvall notes.

The research article, titled "Innocence over utilitarianism: Heightened [moral standards](#) for robots in rescue dilemmas," was published in the *European Journal of Social Psychology*.

  More information: Jukka Sundvall et al, Innocence over utilitarianism: Heightened moral standards for robots in rescue dilemmas, *European Journal of Social Psychology* (2023). [DOI: 10.1002/ejsp.2936](#)

Provided by University of Helsinki