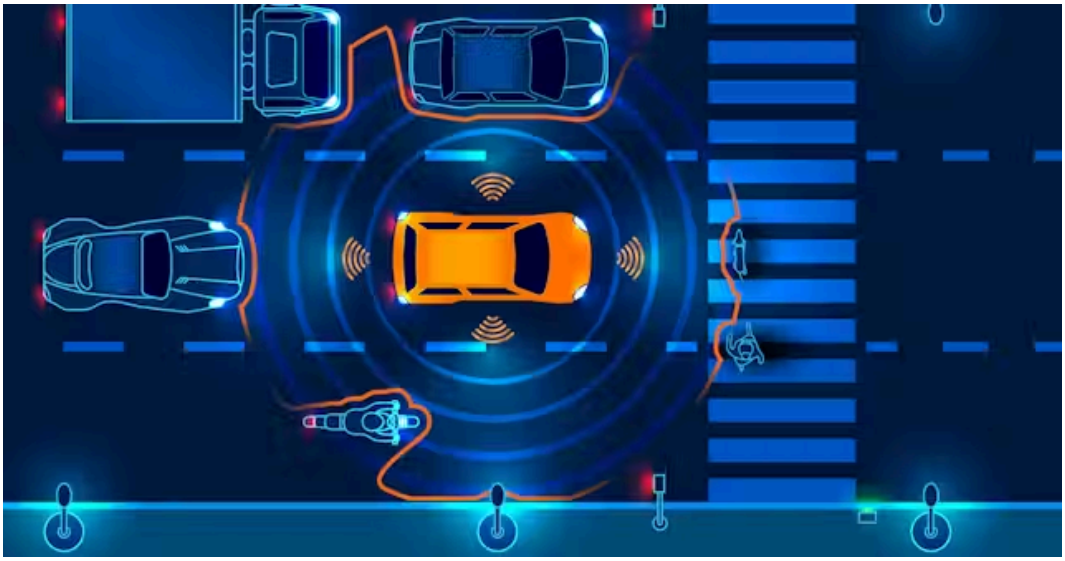
THE CONVERSATION

Academic rigor, journalistic flair



Shutterstock

# The self-driving trolley problem: how will future AI systems make the most ethical choices for all of us?

Published: November 23, 2021 1:56pm EST

**Jumana Abu-Khalaf**

Research Fellow in Computing and Security, Edith Cowan University

**Paul Haskell-Dowland**

Professor of Cyber Security Practice, Edith Cowan University

Artificial intelligence (AI) is already making decisions in the fields of business, health care and manufacturing. But AI algorithms generally still get help from people applying checks and making the final call.

What would happen if AI systems had to make independent decisions, and ones that could mean life or death for humans?

Pop culture has long portrayed our general distrust of AI. In the 2004 sci-fi movie I, Robot, detective Del Spooner (played by Will Smith) is suspicious of robots after being rescued by one from a car crash, while a 12-year-old girl was left to drown. He says:

> *I was the logical choice. It calculated that I had a 45% chance of survival. Sarah only had an 11% chance. That was somebody's baby – 11% is more than enough. A human being would've known that.*

Unlike humans, robots lack a moral conscience and follow the "ethics" programmed into them. At the same time, human morality is highly variable. The "right" thing to do in any situation will depend on who you ask.

For machines to help us to their full potential, we need to make sure they behave ethically. So the question becomes: how do the ethics of AI developers and engineers influence the decisions made by AI?

---

***Read more:*** ***After 75 years, Isaac Asimov's Three Laws of Robotics need updating***

---

## The self-driving future

Imagine a future with self-driving cars that are fully autonomous. If everything works as intended, the morning commute will be an opportunity to prepare for the day's meetings, catch up on news, or sit back and relax.

But what if things go wrong? The car approaches a traffic light, but suddenly the brakes fail and the computer has to make a split-second decision. It can swerve into a nearby pole and kill the passenger, or keep going and kill the pedestrian ahead.

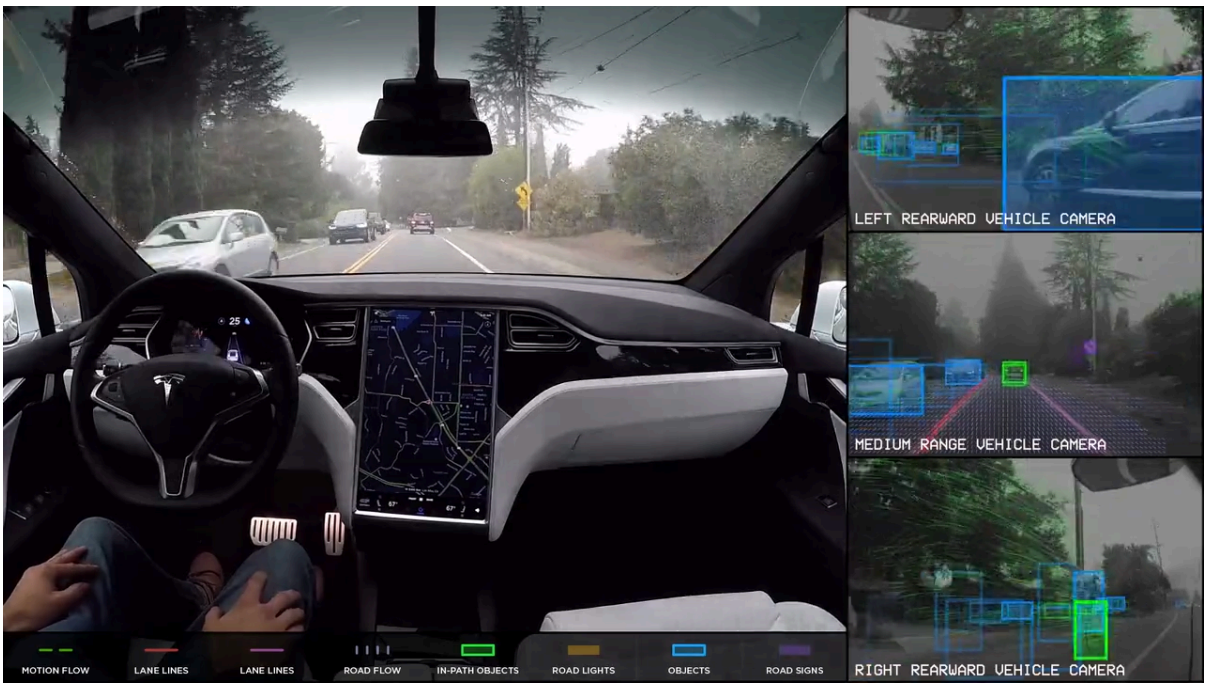
The computer controlling the car will only have access to limited information collected through car sensors, and will have to make a decision based on this. As dramatic as this may seem, we're only a few years away from potentially facing such dilemmas.

Autonomous cars will generally provide safer driving, but accidents will be inevitable – especially in the foreseeable future, when these cars will be sharing the roads with human drivers and other road users.

Tesla does not yet produce fully autonomous cars, although it plans to. In collision situations, Tesla cars don't automatically operate or deactivate the Automatic Emergency Braking (AEB) system if a human driver is in control.

In other words, the driver's actions are not disrupted – even if they themselves are causing the collision. Instead, if the car detects a potential collision, it sends alerts to the driver to take action.

In "autopilot" mode, however, the car should automatically brake for pedestrians. Some argue if the car can prevent a collision, then there is a moral obligation for it to override the driver's actions in every scenario. But would we want an autonomous car to make this decision?




## What's a life worth?

What if a car's computer could evaluate the relative "value" of the passenger in its car and of the pedestrian? If its decision considered this value, technically it would just be making a cost-benefit analysis.

This may sound alarming, but there are already technologies being developed that could allow for this to happen. For instance, the recently re-branded Meta (formerly Facebook) has highly evolved facial recognition that can easily identify individuals in a scene.

---

***Read more:*** ***Facebook will drop its facial recognition system – but here's why we should be sceptical***

---

If these data were incorporated into an autonomous vehicle's AI system, the algorithm could place a dollar value on each life. This possibility is depicted in an extensive 2018 study conducted by experts at the Massachusetts Institute of Technology and colleagues.

Through the Moral Machine experiment, researchers posed various self-driving car scenarios that compelled participants to decide whether to kill a homeless pedestrian or an executive pedestrian.

Results revealed participants' choices depended on the level of economic inequality in their country, wherein more economic inequality meant they were more likely to sacrifice the homeless man.

While not quite as evolved, such data aggregation is already in use with China's social credit system, which decides what social entitlements people have.

The health-care industry is another area where we will see AI making decisions that could save or harm humans. Experts are increasingly developing AI to spot anomalies in medical imaging, and to help physicians in prioritising medical care.

For now, doctors have the final say, but as these technologies become increasingly advanced, what will happen when a doctor and AI algorithm don't make the same diagnosis?

Another example is an automated medicine reminder system. How should the system react if a patient refuses to take their medication? And how does that affect the patient's autonomy, and the overall accountability of the system?

AI-powered drones and weaponry are also ethically concerning, as they can make the decision to kill. There are conflicting views on whether such technologies should be completely banned or regulated. For example, the use of autonomous drones can be limited to surveillance.

Some have called for military robots to be programmed with ethics. But this raises issues about the programmer's accountability in the case where a drone kills civilians by mistake.

---

***Read more:*** ***Gun-toting robo-dogs look like a dystopian nightmare. That's why they offer a powerful moral lesson***

---

## Philosophical dilemmas

There have been many philosophical debates regarding the ethical decisions AI will have to make. The classic example of this is the trolley problem.

People often struggle to make decisions that could have a life-changing outcome. When evaluating how we react to such situations, one study reported choices can vary depending on a range of factors including the respondant's age, gender and culture.

When it comes to AI systems, the algorithms training processes are critical to how they will work in the real world. A system developed in one country can be influenced by the views, politics, ethics and morals of that country, making it unsuitable for use in another place and time.

If the system was controlling aircraft, or guiding a missile, you'd want a high level of confidence it was trained with data that's representative of the environment it's being used in.

Examples of failures and bias in technology implementation have included racist soap dispenser and inappropriate automatic image labelling.

AI is not "good" or "evil". The effects it has on people will depend on the ethics of its developers. So to make the most of it, we'll need to reach a consensus on what we consider "ethical".

While private companies, public organisations and research institutions have their own guidelines for ethical AI, the United Nations has recommended developing what they call "a comprehensive global standard-setting instrument" to provide a global ethical AI framework – and ensure human rights are protected.