Why we need more than just data to create ethical driverless cars

October 25 2018, by Seth Lazar And Colin Klein

Credit: AI-generated image (disclaimer)

What do we want driverless cars to do in unavoidable fatal crashes?

Today researchers published a paper The Moral Machine experiment to address this question.

To create data for the study, almost 40 million people from 233 countries used a website to record decisions about who to save and who to let die in hypothetical driverless car scenarios. It's a version of the classic so-called "trolley dilemma" – where you have to preference people to prioritise in an emergency.

Some of the key findings are intuitive: participants prefer to save people over animals, the young over the old, and more rather than fewer. Other preferences are more troubling: women over men, executives over the homeless, the fit over the obese.

The experiment is unprecedented in both scope and sophistication: we now have a much better sense of how peoples' preferences in such dilemmas vary across the world. The authors, sensibly, caution against taking the results as a simple guide to what self-driving cars should do.

But this is just the first move in what must be a vigorous debate. And in that debate, surveys like these (interesting as they are) can play only a limited role.

How good is our first judgement?

Machines are much faster than us; they don't panic. At their best, they might embody our considered wisdom and apply it efficiently even in harrowing circumstances. To do that, however, we need to start with good data.

Clicks on online quizzes are a great way to find out what people think before they engage their judgment. Yet obviously we don't pander to all prejudices. The authors omitted race and nationality as grounds for choice, and rightly so.

Good survey design can't be done in a vacuum. And moral preferences

are not supposed to just be tastes. To work out the morally right thing to do (think of any morally weighty choice that you have faced), you have to do some serious thinking.

We want to base ethical artificial intelligence on our best judgements, not necessarily our first ones.

The world is 'chancy'

The study used dilemmas that involved two certain outcomes: either you definitely hit the stroller or definitely kill the dog.

But actual decisions involve significant uncertainty: you might be unsure whether the person ahead is a child or a small adult, whether hitting them would kill or injure them, whether a high-speed swerve might work.

Computers might make better predictions, but the world is intrinsically "chancy". This is a big problem. Either-or preferences in certain cases only go so far in telling us what to do in risky ones.

Suppose a self-driving vehicle must choose between letting itself crash and so killing its elderly passenger, or instead veering to the side and killing an infant.

The moral machine experiment predicts that people are on the side of the infant. But it doesn't say by how much we would prefer to spare one over the other. Maybe it's almost a toss-up, and we just lean towards sparing the child. Or maybe saving the child is much more important than saving the pensioner.

Views on this will be extremely diverse, and this survey offers us no guidance. But we can't know how to weigh, say, a 10% probability of killing the child against a 50% probability of killing the pensioner, unless

we know how much more important sparing one is than sparing the other.

Since literally every choice made by driverless cars will be made under uncertainty, this is a significant gap.

What surveys can't tell us

The motivation for the moral machine experiment is understandable. The responsibility of encoding the next generation of ethical artificial intelligence is a daunting one.

Moral disagreement appears rife. A survey looks like a good way to triangulate opinions in a heated world.

But how we handle moral disagreement is not just a scientific problem. It is a moral one too. And, since the times of the ancient Greeks, the solution to that moral problem is not aggregating preferences, but democratic participation.

No doubt democracy is in crisis, at least in parts of the rich world. But it remains our most important tool for making decisions in the presence of unavoidable disagreement.

Democratic decision-making can't be reduced to box ticking. It involves taking your vote seriously, not just clicking a box on a website. It involves participation, debate, and mutual justification.

Surveys like this one cannot tell us why people prefer the options that they do. The fact that a self-driving car's decision correlates with the views of others does not, on its own, justify that choice (imagine a human driver justifying her actions in an accident in the same way).

Mutual justification is the heart of democratic citizenship. And it presupposes engaging not just with what our choices are, but why we make them.

Deciding together

Studies like this are intrinsically interesting, and the authors of this one are admirably explicit about what it is, and what it is not designed to show.

To build on these foundations we need to do much more reflection on how to weigh our moral commitments under uncertainty.

And we need to do so as part of an inclusive democratic process where we don't just aggregate people's preferences, but take seriously the task of deciding, together, our artificial intelligence future.

Provided by The Conversation