Transcript for Iyad Rahawn/TEDxCambridge *What Moral Decisions Should Driverless Cars Make?* https://www.ted.com/talks/iyad_rahwan_what_moral_decisions_should_driverless_cars_make#t-13525

00:10

Today I'm going to talk about technology and society. The Department of Transport estimated that last year 35,000 people died from traffic crashes in the US alone. Worldwide, 1.2 million people die every year in traffic accidents. If there was a way we could eliminate 90 percent of those accidents, would you support it? Of course you would. This is what driverless car technology promises to achieve by eliminating the main source of accidents -- human error.

00:47

Now picture yourself in a driverless car in the year 2030, sitting back and watching this vintage TEDxCambridge video.

00:56

(Laughter)

00:59

All of a sudden, the car experiences mechanical failure and is unable to stop. If the car continues, it will crash into a bunch of pedestrians crossing the street, but the car may swerve, hitting one bystander, killing them to save the pedestrians. What should the car do, and who should decide? What if instead the car could swerve into a wall, crashing and killing you, the passenger, in order to save those pedestrians? This scenario is inspired by the trolley problem, which was invented by philosophers a few decades ago to think about ethics.

01:43

Now, the way we think about this problem matters. We may for example not think about it at all. We may say this scenario is unrealistic, incredibly unlikely, or just silly. But I think this criticism misses the point because it takes the scenario too literally. Of course no accident is going to look like this; no accident has two or three options where everybody dies somehow. Instead, the car is going to calculate something like the probability of hitting a certain group of people, if you swerve one direction versus another direction, you might slightly increase the risk to passengers or other drivers versus pedestrians. It's going to be a more complex calculation, but it's still going to involve trade-offs, and trade-offs often require ethics.

02:37

We might say then, "Well, let's not worry about this. Let's wait until technology is fully ready and 100 percent safe." Suppose that we can indeed eliminate 90 percent of those accidents, or even 99 percent in the next 10 years. What if eliminating the last one percent of accidents requires 50 more years of research? Should we not adopt the technology? That's 60 million people dead in car accidents if we maintain the current rate. So the point is, waiting for full safety is also a choice, and it also involves trade-offs.

03:21

People online on social media have been coming up with all sorts of ways to not think about this problem. One person suggested the car should just swerve somehow in between the passengers --

03:32

(Laughter)

03:33

and the bystander. Of course if that's what the car can do, that's what the car should do. We're interested in scenarios in which this is not possible. And my personal favorite was a suggestion by a blogger to have an eject button in the car that you press --

03:51

(Laughter)

03:52

just before the car self-destructs.

03:54

(Laughter)

03:57

So if we acknowledge that cars will have to make trade-offs on the road, how do we think about those trade-offs, and how do we decide? Well, maybe we should run a survey to find out what society wants, because ultimately, regulations and the law are a reflection of societal values.

04:17

So this is what we did. With my collaborators, Jean-François Bonnefon and Azim Shariff, we ran a survey in which we presented people with these types of scenarios. We gave them two options inspired by two philosophers: Jeremy Bentham and Immanuel Kant. Bentham says the car should follow utilitarian ethics: it should take the action that will minimize total harm -- even if that action will kill a bystander and even if that action will kill the passenger. Immanuel Kant says the car should follow duty-bound principles, like "Thou shalt not kill." So you should not take an action that explicitly harms a human being, and you should let the car take its course even if that's going to harm more people.

05:05

What do you think? Bentham or Kant? Here's what we found. Most people sided with Bentham. So it seems that people want cars to be utilitarian, minimize total harm, and that's what we should all do. Problem solved. But there is a little catch. When we asked people whether they would purchase such cars, they said, "Absolutely not."

05:30

(Laughter)

05:33

They would like to buy cars that protect them at all costs, but they want everybody else to buy cars that minimize harm.

05:40

(Laughter)

05:44

We've seen this problem before. It's called a social dilemma. And to understand the social dilemma, we have to go a little bit back in history. In the 1800s, English economist William Forster Lloyd published a pamphlet which describes the following scenario. You have a group of farmers -- English farmers -- who are sharing a common land for their sheep to graze. Now, if each farmer brings a certain number of sheep -- let's say three sheep -- the land will be rejuvenated, the farmers are happy, the sheep are happy, everything is good. Now, if one farmer brings one extra sheep, that farmer will do slightly better, and no one else will be harmed. But if every farmer made that individually rational decision, the land will be overrun, and it will be depleted to the detriment of all the farmers, and of course, to the detriment of the sheep.

06:42

We see this problem in many places: in the difficulty of managing overfishing, or in reducing carbon emissions to mitigate climate change. When it comes to the regulation of driverless cars, the common land now is basically public safety -- that's the common good -- and the farmers are the passengers or the car owners who are choosing to ride in those cars. And by making the individually rational choice of prioritizing their own safety, they may collectively be diminishing the common good, which is minimizing total harm. It's called the tragedy of the commons, traditionally, but I think in the case of driverless cars, the problem may be a little bit more insidious because there is not necessarily an individual human being making those decisions. So car manufacturers may simply program cars that will maximize safety for their clients, and those cars may learn automatically on their own that doing so requires slightly increasing risk for pedestrians. So to use the sheep metaphor, it's like we now have electric sheep that have a mind of their own.

08:02

(Laughter)

08:03

And they may go and graze even if the farmer doesn't know it.

08:08

So this is what we may call the tragedy of the algorithmic commons, and if offers new types of challenges. Typically, traditionally, we solve these types of social dilemmas using regulation, so either governments or communities get together, and they decide collectively what kind of outcome they want and what sort of constraints on individual behavior they need to implement. And then using monitoring and enforcement, they can make sure that the public good is preserved. So why don't we just, as regulators, require that all cars minimize harm? After all, this is what people say they want. And more importantly, I can be sure that as an individual, if I buy a car that may sacrifice me in a very rare case, I'm not the only sucker doing that while everybody else enjoys unconditional protection.

09:06

In our survey, we did ask people whether they would support regulation and here's what we found. First of all, people said no to regulation; and second, they said, "Well if you regulate cars to do this and to minimize total harm, I will not buy those cars." So ironically, by regulating cars to minimize harm, we may actually end up with more harm because people may not opt into the safer technology even if it's much safer than human drivers.

09:39

I don't have the final answer to this riddle, but I think as a starting point, we need society to come together to decide what trade-offs we are comfortable with and to come up with ways in which we can enforce those trade-offs.

09:56

As a starting point, my brilliant students, Edmond Awad and Sohan Dsouza, built the Moral Machine website, which generates random scenarios at you -- basically a bunch of random dilemmas in a sequence where you have to choose what the car should do in a given scenario. And we vary the ages and even the species of the different victims. So far we've collected over five million decisions by over one million people worldwide from the website. And this is helping us form an early picture of what trade-offs people are comfortable with and what matters to them -- even across cultures. But more importantly, doing this exercise is helping people recognize the difficulty of making those choices and that the regulators are tasked with impossible choices. And maybe this will help us as a society understand the kinds of trade-offs that will be implemented ultimately in regulation.

10:59

And indeed, I was very happy to hear that the first set of regulations that came from the Department of Transport -- announced last week -- included a 15-point checklist for all carmakers to provide, and number 14 was ethical consideration -- how are you going to deal with that. We also have people reflect on their own decisions by giving them summaries of what they chose. I'll give you one example -- I'm just going to warn you that this is not your typical example, your typical user. This is the most sacrificed and the most saved character for this person.

11:38

(Laughter)

11:44

Some of you may agree with him, or her, we don't know. But this person also seems to slightly prefer passengers over pedestrians in their choices and is very happy to punish jaywalking.

12:01

(Laughter)

12:06

So let's wrap up. We started with the question -- let's call it the ethical dilemma -- of what the car should do in a specific scenario: swerve or stay? But then we realized that the problem was a different one. It was the problem of how to get society to agree on and enforce the trade-offs they're comfortable with. It's a social dilemma.

12:27

In the 1940s, Isaac Asimov wrote his famous laws of robotics -- the three laws of robotics. A robot may not harm a human being, a robot may not disobey a human being, and a robot may not allow itself to come to harm -- in this order of importance. But after 40 years or so and after so many stories pushing these laws to the limit, Asimov introduced the zeroth law which takes precedence above all, and it's that a robot may not harm humanity as a whole. I don't know what this means in the context of driverless cars or any specific situation, and I don't know how we can implement it, but I think that by recognizing that the regulation of driverless cars is not only a technological problem but also a societal cooperation problem, I hope that we can at least begin to ask the right questions.

13:26

Thank you.

13:27

(Applause)