ACM.org | Join ACM | About Communications | ACM Resources | Alerts & Feeds

SIGN OUT (Miltiades Anagnostou)

Search

HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RESEARCH | PRACTICE | CAREERS | ARCHIVE | VIDEOS

Home / Blogs / BLOG@CACM / Is the Trolley Problem Useful for Studying Autonomous... / Full Text

BLOG@CACM

Is the Trolley Problem Useful for Studying Autonomous Vehicles?

By Jason Hong
May 2, 2019
Comments

PRINT

VIEW AS: SHARE:

Does the trolley problem offer any useful insights for autonomous vehicles, in terms of design of or public policy around these systems?

The trolley problem, if you're not familiar with it, asks people to consider hypothetical situations where a trolley will crash into one of two groups. An individual has to make a decision about which group the trolley will crash into and kill. A typical scenario is to choose between the trolley crashing into and killing a single individual or a group of five people.

The trolley problem has seen a resurgence because of its natural connection with autonomous vehicles. For example, an autonomous vehicle is going too fast and can't stop in time, should it be programmed to crash into a person or an animal? The elderly man or the young child? The pregnant woman or two individuals? Let’s call this the Autonomous Trolley problem.

The Autonomous Trolley problem is a lot like Asimov's Three Laws of Robotics. It's easy to understand and fun for philosophers and journalists to talk about. It also gives little practical insight for developers and policy makers, while also sucking the oxygen out of the room for more pressing discussions about societal values that should be embodied in autonomous vehicles and the inevitable tradeoffs those values will entail.

One issue with the Autonomous Trolley problem is that it makes a lot of unrealistic assumptions that would rarely ever happen in the real world. Most notably, the problems are framed such that the autonomous vehicle is the only entity with any agency. Or to put it more simply, pedestrians are completely passive and won't try to dodge oncoming vehicles.

Another issue with the Autonomous Trolley problem is that it assumes that the car is an oracle with perfect sensing and completely accurate predictive modeling. The vehicle can use its sensors to differentiate between animals, children, pregnant women, and jaywalkers without any errors. Furthermore, it can correctly predict exactly how many individuals will die if a decision is made. If autonomous vehicles really did have such capabilities, it would be much better to use that to slow down earlier and not hit anything.

Perhaps the most important issue with the Autonomous Trolley problem is that it’s not clear what insights car manufacturers should take from results of these experiments. Driving is often a fluid situation with multiple independent actors, and in many cases, if an autonomous vehicle is about to crash into something, the primary failure likely happened several seconds previously. It seems far more useful for product teams to devote energy to preventing these problems, rather than implementing a "who should die" algorithm that would not only be impractical due to the fact that oracles don’t exist, but also because of the liability that such an algorithm would entail.

MORE NEWS & OPINIONS

All Hail the AI Overlord: Smart Cities and the AI Internet of Things
Ars Technica

What Trump's Executive Order on AI Is Missing
Wired

The Artificialistic Fallacy
Robin K. Hill

ACM RESOURCES

Working with Flash 5 and Generator
Courses

As an alternative to studying Autonomous Trolley problems, I'd suggest researchers look more at how to elicit the values that different stakeholders have—for example, passengers of autonomous vehicles, other drivers, pedestrians, and society as a whole—and how best to balance these values.

For example, Rodney Brooks describes a scenario where, rather than parking their autonomous vehicle, people might just have their car go up and down the street until they are ready to leave. This kind of behavior might be advantageous for the people using the autonomous vehicle, but bad for everyone else. Should these kinds of behaviors be allowed? If so, to what degree? Would circling around the block just once or twice be acceptable? Also, if these kinds of behaviors are generally deemed bad, how to actually limit them? Should it be through some kind of technical mechanism (e.g. cars can only be passenger-less for at most five minutes), some kind of social mechanism (e.g. the car has a display on its outside saying "I am going to pick up my owner soon"), or perhaps some kind of oversight by police?

Here's another set of scenarios: how should autonomous vehicles operate with respect to social norms that may have arisen in a community? For example, I've met some people who were really, really possessive about the street parking in front of their houses, that it's "their spot" and they don't appreciate anyone else parking there (no, I've never had awkward experiences with people like this, why do you ask?). Should these citizens be able to post some kind of marker on the street denoting that it's "their spot"? What about people who have legitimate accessibility needs?

As another example, in Pittsburgh, we sometimes get really heavy snowfall in winter. People who dig their cars out of the snow sometimes leave chairs to mark the parking spot as theirs. However, this also means that the parking spot is underutilized. Would it be fair to allow other autonomous vehicles to use that parking spot, as long as the car is gone before the spot's "owner" is back?

What about street cleaning? In our city, specialized vehicles periodically sweep away leaves on the street, for example sweeping one side of the street on the first Tuesday of the month and the other side on the first Wednesday. However, sometimes people forget to move their cars and get a parking fine. Should there be some mechanism for other people to temporarily move other people's cars? Should parking fines exist at all if police can request a car to temporarily move?

Stepping back, parking is a limited resource, but autonomous vehicles offer us the opportunity to re-think how parking is done. Perhaps each neighborhood should have centralized parking for autonomous vehicles, and people can just summon their cars. Perhaps street parking should have the illusion of being reserved for "owners", with autonomous vehicles shuffling out of the way just in time. If you can summon your car to you, does it really matter to you where you parked? Perhaps a residential neighborhood might want to ban parked cars completely.

Let's move on to driving behaviors. Should autonomous vehicles aim to match prescriptive driving behaviors (i.e. always obey the speed limit) or descriptive norms of how people actually drive? For example, in 2007, some students in created a video of themselves driving exactly the speed limit on the Interstate 285 beltway in Atlanta (55 mph at the time). However, in practice, people typically drive 75 mph or faster on this highway. Their video showed a huge backup in traffic, with people even using the shoulder of the road to get around them. What happens when speed limits and actual driving behaviors don't match? What should autonomous vehicles do?

On a similar note, should passengers be able to override an autonomous vehicle's behaviors in some circumstances? For example, what if you are stuck at a red light that hasn't changed in 10 minutes, should there be an option to tell the car to just go? What happens if there is a police officer waving you forward, but the autonomous vehicle hasn't registered that correctly? What if you are stuck in highway traffic but your exit is just five cars ahead, should you be able to tell your vehicle to just go on the shoulder and get off? If so, how should these options be presented to passengers? Should the passenger's decision be logged in some way? Should the passenger's actions be reviewed in some way? If these mechanisms do exist, how to avoid inevitable abuse by people?

It's also likely that autonomous vehicles will have an array of cameras. People have already recorded a wide range of things on existing dashboard cameras, including car crashes, street fights, and even a meteor in Russia. If your GPS shows that you were in the vicinity of an incident (e.g. a car accident or a pedestrian getting hit), should the police or the victims be notified that that video exists? It's also possible to network video streams for all of these autonomous vehicles together and apply computer vision algorithms, essentially getting a city-wide live feed. Should such a thing be allowed? What if it were used to find empty parking spots? Or to measure number of pedestrians in a neighborhood? Recording the location of license plates? Help city planners understand how people use the city?

Today, drivers use a set of techniques to negotiate with other drivers, bicyclists, and pedestrians as to their intentions, for example using turn signals or using one's hand to wave people forward. What should the equivalent be for autonomous vehicles? Similarly, how can we make autonomous vehicle behaviors more predictable to others? Should there be external screens with a standard set of driving symbols to give everyone around the vehicle some kind of feedback?

Should autonomous vehicles honk at drivers of non-autonomous vehicles that have not noticed that the light has changed several seconds ago? Should autonomous vehicles always yield to pedestrians and bicyclists? Are there ways for pedestrians and bicyclists be able to give signals as to their intentions? For example, sometimes pedestrians will wave, signaling to the car to go first. What kind of feedback can the autonomous vehicle give to acknowledge the signal? What happens if the pedestrian is deliberately trolling the autonomous vehicle, just pretending to cross the street so as to stop the car [Brooks]? What happens if the pedestrian is blind?

What about making things safer for pedestrians and bicyclists? Many cities implicitly prioritize cars over other forms of transportation. If a city or a society decides it wants to prioritize pedestrians and bicyclists, how might an autonomous vehicle be designed for these values?

While I'm personally skeptical that we'll see Level 4 or Level 5 autonomous vehicles this coming decade, I think that they will be common within a generation. That's good, this gives us time to actually work through the difficult and thorny questions surrounding autonomous vehicles. These questions aren't as fun to consider as the Autonomous Trolley problem. There's also not necessarily a right answer for these questions, it's primarily a question of understanding the different values that different stakeholders have, what the tradeoffs are, and which ones should take precedence, while also being grounded in what is practical and feasible with technology. Just as automobiles fundamentally reshaped society in the 20th century (think highways, traffic jams, gas stations, smog, freedom for teenagers, car chases in movies, and more), these autonomous vehicles have the potential to do the same in the 21st. Let's make sure we design these autonomous vehicles to be part of a world that we would all want to live in.

No entries found

Comment on this article

Signed comments submitted to this site are moderated and will appear if they are relevant to the topic and not abusive. Your comment will appear with your username if published. View our policy on comments

☐ Notify me via email when subsequent user comments are published with this article.