AI's "Trolley Problem" Problem

**Wednesday 16 Aug 2017**

**[*Image by McGeddon (CC BY-SA 4.0)*](https://commons.wikimedia.org/w/index.php?curid=52237245)

The striking ascent of self-driving cars, from the stuff of sci-fi to a dealership near you, offers one of the most transformative examples of the impact of artificial intelligence on society. Cars are a technology that almost everyone in developed societies will use at some point in their lives (whether as driver or passenger), and so the prospect that everything from pick-up to pull-up will be fully automated will likely prove one of the most fundamental transformations of this adolescent century.

The adoption of earlier inventions like email and smartphones was hastened by their seeming similarity to existing technologies like (snail) mail and (analogue) phones, and the fact that self-driving cars seem to be simply “cars that can drive themselves” will no doubt serve a similar purpose. (This does leave open the question of how we’ll describe conventional cars in future: “pedal cars”?)

Yet just as with email and smartphones, looking “under the hood” of self-driving cars reveals a much more sophisticated set of functions and potential abilities than their conventional forebears. Yes, self-driving cars will drive themselves, but they’ll likely be able to do so much more as well. They might “intelligently” decide when to arrive at your house for your morning commute, based on the contents of your calendar that day. They might know (from your fridge — or your glucose monitor) which groceries you’re low on, and autonomously go fetch some (just in case the thirty minute wait for an Amazon drone proves too onerous). In a similar sense, even the first iteration of the iPhone — which didn’t yet have an App Store — nonetheless had extraordinary potential as a multimedia device due in part to its connectivity.

As well as these exciting new possibilities, however, self-driving cars will also have to go out of their way *not* to do something: kill people. [**35,000 people died on American roads last year**](http://fortune.com/2016/09/14/distracted-driving-epidemic/) — probably more than those killed by guns, and certainly more than those killed by phones of any sort (though cars and smartphones have proved an especially dangerous combination.)

The attempts to fully automate such a lethal technology have given not only inventors but also regulators, academics and journalists much to ponder, to a far greater extent than with earlier consumer technology breakthroughs. By far, the question receiving the most prominent discussion is the so-called “trolley problem”. This thought experiment is a longstanding ethical paradox. [**Borrowing Wikipedia’s summary**](https://en.wikipedia.org/wiki/trolley_problem), the problem states:

**There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are five people tied up and unable to move. The trolley is headed straight for them. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is one person on the side track. You have two options:**

1. **Do nothing, and the trolley kills the five people on the main track.**
2. **Pull the lever, diverting the trolley onto the side track where it will kill one person.**

**Which is the most ethical choice?**

This problem gets to the heart of some of the oldest debates in moral philosophy, not least the divide between consequentialist and utilitarian approaches — which seek to optimise the “greatest good for the greatest number” and emphasise the *impact* of someone’s actions — and deontological ethics, which hold that participation in some action might always be wrong, with proportionately less regard to the consequences.

The problem has been extended to include a related case in which, instead of flipping a switch, the only way to save the five people is to push a nearby man off a bridge (since he is large enough to stop the trolley in its tracks). Research suggests that this alternative scenario causes many people to change their mind: many people are comfortable flipping the switch but not shoving the man. Though this implies a moral distinction between these two acts (despite their identical effect: sacrificing one life to save five), [**research suggests**](http://science.sciencemag.org/content/293/5537/2105) that the divergent attitudes ultimately result from a neurological distinction, as different parts of subjects’ brains were observed as controlling the decisions in the different cases.

**Creating a moral motorist**

In many ways it is easy to see why the trolley problem has become the canonical example in thinking about self-driving cars. Given the death toll that human-operated cars deliver every year, it is safe to assume that cars will continue to be at least somewhat hazardous, even with the vast improvements in efficiency that automation might bring. It follows, then, that we should think about the road-based equivalent of this track-based trolley problem as a matter of urgency, deciding whether and how to code societal values into autonomous vehicles.

Indeed, scientists at MIT’s Media Lab have launched an impressive attempt to experiment with just this question. The [**Moral Machine platform**](http://moralmachine.mit.edu/) invites users to judge a series of hypothetical scenarios, making difficult decisions about the direction an out-of-control car should swerve. After answering a series of questions, the survey will rank a user’s implied “preferences” with almost disturbing granularity, in terms of gender, age, wealth, health, and much else.

The Moral Machine effort is laudable, and its significance self-evident. It serves to underline that the rise of automated technology, and specifically artificial intelligence, may have the unintended positive effect of encouraging society to be more open and explicit about those values. Code — whether in the form of law, or technological architecture — necessitates clarity, at least of a sort. Enabling a machine to make decisions demands declarations of our more fundamental values on which those decisions should rest.

And yet, the prevalence of the trolley problem in the context of self-driving cars is itself, in a sense, problematic. At the most basic level, we are talking about self-driving cars on roads, not trolleys on a track*,* which opens up a far wider array of possibilities for aversive action — including skidding in any direction, rather than just taking one course or another, as in the trolley example.

Since the trolley problem is purely hypothetical, it can be easily adapted for other contexts — the Moral Machine project, of course, adapts it for the road. But the basic simplicity that makes the problem so popular — the trolley or car can swerve left or right , so our decision is both binary and binding — is also what makes it problematic, in terms of reasoning about AI.

By asking ourselves what *we* would do when faced with such an ethically thorny issue, we risk ascribing to AI a “thought process” that it doesn’t really have. The trolley problem is an ethical paradox, which forces us reflect on our own values and biases. Though the fictitious problem involves the subject making a quick decision, the exercise is useful *precisely because* it shows how hard making such a decision would be in practice. Paradoxes that are easily solvable are not worthy of the name.

Machines are less prone to introspection. A self-driving car would be able to execute a “decision” in milliseconds, but its decision-making process is unlikely to operate much like our own. True, self-driving cars take on a wealth of data from the surrounding argument using cameras and radar much as human drivers do using eyes and ears, and true, neural networks — which are designed to mimic the human brain — can be used to help recognise objects and [**even predict pedestrians’ movement**](http://spectrum.ieee.org/computing/embedded-systems/bringing-big-neural-networks-to-selfdriving-cars-smartphones-and-drones).

But these similarities are analogical, not biological. To the extent that an ethical preference (e.g., swerve into the dog to avoid the group of schoolchildren) can be coded into self-driving cars, the decision-making process would not be “rational” in a way we would understand. No ethical considerations would be in play *during the making of the decision*; rather, the decision would result from a set of *pre-existing preferences* implemented by coders — perhaps informed, as in the Moral Machine project, by the wider public.

This is not to say that a decision with ethical consequences which is processed by a machine would necessarily be worse than one made by a human in a typical crash situation. In fact, it might well be better: the self-driving car would be drawing on pre-existing preferences to swerve this way or that, and so hypothetically, if these preferences were developed equitably, ethically, and with the input from a representative group of people, they might be more reflective of a collective human will than a split-second decision taken by an individual person, “programmed” with all her specific biases and particular life experiences.

This is almost certainly naive, as technologies very often reflect the intentions and even the characteristics of their designers: consider, as an extreme example, facial recognition software, which [**can fail to recognise darker faces as effectively as lighter ones**](https://www.recode.net/2017/1/18/14304964/data-facial-recognition-trouble-recognizing-black-white-faces-diversity).

Regardless of the answer, the question is the wrong one to ask.

If, [**as the available data suggest**](https://www.wired.com/2017/02/california-dmv-autonomous-car-disengagement/), self-driving cars are much less likely to be involved in fatal accidents, only the steeliest utilitarian would argue that means we should worry less about these situations. But it *does* suggest that the trolley problem is the wrong lens through which to think about self-driving cars and the artificial intelligence which will power them. Not only does the trolley problem give the wrong impression of how automated decision-making works, but it limits our focus to a very specific set of hypothetical possibilities, potentially distracting us from far weightier challenges.

**Road rage against the machine**

The most important of these is the possibility of deliberate manipulation. At a time when organisations ranging from banks to political parties to cable TV networks can be hacked seemingly at will, the potential for foul play in the software used to safely get people from A to B seems alarmingly plausible. Alluding to the fact that these hacks tend to have a financial motivation attached, the influential academic and writer [**Zeynep Tufekci**](https://medium.com/u/a5b491a8b18c) recently posed a far more pertinent and problematic question:

In contrast to the trolley problem, Tufekci’s hypothetical far better reflects the actual, near-future dangers of self-driving car technology. Though on occasions self-driving cars may face versions of the trolley problem when in the wild, the far greater danger is that their systems might be wilfully hacked by greedy humans.

Nor does hacking need to target self-driving systems directly. [**A new paper by researchers from four American universities**](https://arxiv.org/pdf/1707.08945.pdf) shows that self-driving cars can be quite trivially manipulated by their surrounding environment. By subtly altering road signs using stickers, which might look like random vandalism, the researchers were able to fool autonomous vehicles into erroneously perceiving a stop sign as a speed limit sign.

It’s easy to see how the consequences of these sorts of attacks could be devastating, without the need to even target the vehicle’s software directly. While specific fixes will surely be developed, these problems — we might call them the “cliff-edge problem” and the “stop sign problem” — are in reality far more dangerous to the safe operation of self-driving cars than the traditional trolley problem.

As a technology, self-driving cars present enormous opportunity for safe, productive travel. The fact that they are, on aggregate, much safer than human-operated vehicles shouldn’t preclude discussions about how to “teach” such machines to handle ethically complicated situations, including the values they should draw on when these arise. (And these reflections might prove as useful for our own ethical conduct as for that of the cars themselves.)

But in contrast to consumer devices like the telephone, the technology that self-driving car designers hope to transform is already extremely dangerous — due in large part to everyday negligence and incompetence on the part of human drivers. As the “cliff-edge” and “stop sign” problems demonstrate, the core danger of self-driving cars is not what happens when we improve on human incompetence, but instead what happens when we enable human malevolence.

*Josh Cowls is a Research Assistant in Data Ethics at the Alan Turing Institute. This post was originally published*[***on his Medium blog***](https://medium.com/%40joshcowls/ai-and-the-trolley-problem-problem-ef48582b49bf)*. You can follow Josh on Twitter*[***@JoshCowls***](https://twitter.com/joshcowls)*.*



[**Josh Cowls**](https://www.turing.ac.uk/people/researchers/josh-cowls)

Research Associate and Doctoral Student

<https://www.turing.ac.uk/blog/ais-trolley-problem-problem>





# 2021 Disengagement Report from California

[Mario Herger](https://thelastdriverlicenseholder.com/author/mherger/)

[February 9, 2022](https://thelastdriverlicenseholder.com/2022/02/09/2021-disengagement-report-from-california/)

[Autonomous Car](https://thelastdriverlicenseholder.com/category/autonomous-car/)

## Post navigation



The beginning of February is the moment when the California DMV publishes its annual Disengagement Report, or more precisely, the associated [data](https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/disengagement-reports/). A disengagement report is required from the nearly 50 companies currently licensed to test autonomous cars on public roads in California. This has been in place for a number of years, I have documented [2015, 2016, 2017](https://thelastdriverlicenseholder.com/2018/02/01/disengagement-report-2017-the-good-the-bad-the-ugly/), [2018](https://thelastdriverlicenseholder.com/2019/02/13/update-disengagement-reports-2018-final-results/), [2019](https://thelastdriverlicenseholder.com/2020/02/26/disengagement-report-2019/) and [2020](https://thelastdriverlicenseholder.com/2021/02/09/2020-disengagement-reports-from-california/).

Each year, licensees must report how many autonomous test and commercial vehicles they operated from Dec. 1 to Nov. 30, how many miles or kilometers they drove, and how many disengagements there were. A disengagement here can be described as the event when the autonomous vehicle either got stuck and handed over control to a safety driver present in the vehicle, or when a safety driver took control himself. The latter is the case if, for example, the vehicle was about to make a mistake or could no longer find its way out of a traffic situation.

One can already see from this definition that it is at the discretion of the licensees what is reported as a disengagement. Often an overzealous safety driver takes control, where it then becomes apparent in the simulation that the vehicle would have handled that situation itself after all. Also, fewer interventions are expected on freeways than in suburbs or in cities like San Francisco, where many different road users congregate. In this respect, the figures reported and published by the companies should be taken with a grain of salt. However, I interpret and present them anyway because California is currently the only region in the world where we have such data publicly available, and they offer a little insight into the progress and state of development of autonomous vehicles.

A total of 26 companies reported data for the full year ( (December 1st, 2020 to November 30th, 2021)), as did two others (Inceptio and Uber), but Uber (UATC) was sold to Aurora and thus no longer appears as separate entity. Another company (Imagry) did not report and as a result lost its license. Some other companies either did not report any activities or did not have to submit a report, therefore 26 companies are represented in the following analysis. Furthermore, 7 manufacturers currently have the license to operate the vehicles even without a safety driver, an eighth company (Pony.AI) had lost this license in the fall of 2021.

### Reports with Safety Drivers

First up are the reports of rides with safety drivers in the vehicle. A total of 1,175 autonomous vehicles had been in operation in California during the period. Waymo had the most vehicles in operation with 693, followed by Cruise with 138.



Number of vehicles per manufacturer

All manufacturers together covered a total of 6,482,960 kilometers (4,051,850 miles). Waymo achieved the most kilometers with 3,721,349 kilometers, followed by Cruise with 1,401,768 kilometers.



Number of miles driven per manufacturer

The number of disengagements varies among manufacturers, but the calculation using the number of kilometers driven shows an interesting picture. For example, Waymo reported one disengagement per 12,744 kilometers for 2021, which is now four times more common than in 2020, when the company reported 47,911 kilometers per disengagement. This likely has to do with the fact that Waymo has now shifted the focus of driving around its Mountain View headquarters entirely to San Francisco. So all the vehicles in Mountain View and the surrounding area are gone (they [report](https://thelastdriverlicenseholder.com/2021/12/10/where-is-waymos-mountain-view-fleet-gone/)), but at least two [depots](https://thelastdriverlicenseholder.com/2021/11/02/waymos-san-francisco-depot-now-with-117-vehicles-and-36-charging-points/) have been opened in San Francisco. And San Francisco is a much more complex place for autonomous vehicles (and human drivers) than Mountain View.

Another reason for Waymo’s more frequent interventions is also said to be the switch to a new type of vehicle. While Chrysler Pacifica minivans were previously the workhorse of the fleet, it is now the all-electric Jaguar iPace. There seem to have been a bit more challenges than probably expected.



Number of miles driven per disengagement

Cruise, a division of GM, in turn, has been developing its vehicles in San Francisco from the beginning. The disengagement rate of once per 66,751 kilometers seems to indicate this. This was down almost 50 percent from the previous year’s 45,632 kilometers. To that end, China-based startup AutoX came out on top with 80,173 kilometers per disengagement. However, the company drove only one 46th of Waymo’s distance and reported only one disengagement at all. A similar story can be seen at Argo.AI. There, too, only one disengagement was reported.

### A screenshot of a computer  Description automatically generated

### Driverless Reports

Four companies (Apollo, Cruise, Nuro, and Pony.AI) reported driverless trips. In total, these four companies reeled off 39,972 kilometers during the comparison period.

Pony.AI had an [incident](https://thelastdriverlicenseholder.com/2021/12/10/pony-ai-no-longer-has-a-driverless-license/) that caused the DMV to revoke the company’s license. However, we can see that impressive distances have already been covered by driverless rides.

### More details

Again, Tesla does not appear this year, and the reason lies in the definition of what is considered an autonomous test vehicle. Tesla’s current interpretation of the FSD Beta, which is now in use in about 60,000 customer vehicles, is that it is a driver assistance system that does not require reporting to the DMV. California is currently working to close this interpretation gap.

Apple, on the other hand, does not seem to be a competitor to be reckoned with in the near future with the current state of development; the numbers are too far away from the top companies. It is interesting to see that many Chinese companies are in the forefront and are putting significant effort into it. AutoX, for example, now has [1,000 vehicles](https://cnevpost.com/2022/02/09/alibaba-backed-autoxs-robotaxi-fleet-reaches-1000-vehicles/) in operation in China.

### Conclusion

We see a densification of the field, increasing mileage per disengagement, but also challenges that rides in new cities and regions and with new vehicles offer to companies. We are particularly pleased to see driverless rides included for the first time. Even though it was “only” 40,000 miles this time, a big jump is expected with next year, as Waymo is about to launch its driverless robotaxi operation in San Francisco, and Cruise will presumably be able to expand driverless operations, previously limited to between 11 p.m. and 5 a.m.

<https://thelastdriverlicenseholder.com/2022/02/09/2021-disengagement-report-from-california/>