Why Everyone Has It Wrong About the Ethics of Autonomous Vehicles

John Basl, Northeastern University

Jeff Behrends, Harvard University

Introduction

Autonomous vehicles (AVs) raise a host of ethical challenges, including determining how AVs should interact with human drivers in mixed-traffic environments, assigning responsibility when AVs crash or cause a crash, how to manage the social and economic impact of AVs that displace human workers, etc. However, public and academic discussion of the ethics of AVs has been dominated by the question of how to program AVs to manage accident scenarios, and in particular whether and how to draw on so-called "trolley cases" to help us resolve this issue. Some in the debate are optimistic that trolley cases will be especially useful when addressing accident scenarios, while others are pessimistic, insisting that trolley cases are of little to no value.

In this paper we summarize the extant debate between the optimists and pessimists, articulate why both sides of the debate have failed to recognize the appropriate relationship between trolley cases and AV design, and explain how to better draw on the resources of philosophy to resolve issues in the ethics of AV design and development.

Trolley Optimism

AVs will inevitably be in *accident scenarios*, scenarios where an accident that will cause harm (to pedestrians, passengers, etc.) is unavoidable. Whereas human drivers in these circumstances have very limited ability to navigate them with any sort of control, AVs might be in a position to "decide" how to distribute those harms. It has seemed to many that because with AVs there is some ability to exercise control over how harms are distributed, we must think carefully about how to program AVs for accident scenarios. The question is how should we do so?

It has not escaped notice that some accident scenarios bear a resemblance to what are known in philosophy as “trolley cases”. A trolley case is an imagined scenario in which a runaway trolley will continue on its course resulting in the death of some number of individuals unless some choice is made to divert or otherwise alter the course of the trolley, resulting in some other number of deaths. In the most classic trolley case, the trolley is headed down a track and will kill five people that can’t escape. A bystander has the ability to pull a switch, diverting the trolley onto another track. However, on this track there is one person who cannot escape and will die if the trolley is diverted. We can imagine an AV that is traveling down a street when suddenly a group of pedestrians runs into the street. The only way to avoid hitting them is to take a turn that will result in the death of a pedestrian on the sidewalk. In another version of the trolley case, a trolley cannot stop and will kill five people unless an object of sufficient weight is pushed in front of the trolley. A bystander has the option of pushing a large person off a bridge and onto the tracks in a way that would stop the train before it kills the five. Again, we can imagine a case involving an AV that has a similar structure: Perhaps there is an empty AV that has gone out of control and will hit five pedestrians unless another AV with a single passenger in it drives itself into the first AV.

Trolley Optimism is the view that we can and should draw on the resources of trolley cases to inform how we should program AVs to behave in these sorts of accident scenarios. The general proposal is that we can construct trolley cases of various kinds, reach a verdict about what action or behavior is appropriate in that case, and then apply that verdict in the case of AVs, programming vehicles to behave in a way that mirrors the correct decision in the analogous trolley case (Wallach and Allen 2009; Lin 2013; Hübner and White 2018).

While trolley cases may be borne of philosophy, Trolley Optimism is not confined to philosophy departments (see Worstall 2014; Achenbach 2015; Doctorow 2015; Hao 2018). Consider MIT's Moral Machine project which has a variety of components. One component is a website which presents visitors with a variety of accident scenarios, asking about each how the visitor thinks the car ought to behave in that scenario. These scenarios involve many variables, testing visitors' judgments about, for example, how to trade off people and animals, men and women, the elderly and children, those that obey walk signals and those that don't, etc. While some might see the Moral Machine project as simply a tool for collecting sociological data, others think that this data, in aggregate, should be used to decide how AVs should be programmed to behave in accident scenarios (Noothigattu et al. 2017). Whereas philosophers might endorse a variety of Trolley Optimism on which we fight it out and figure out the correct thing to do in a trolley case, using that to tell us how to make AVs behave in accident scenarios, this democratic variant of Trolley Optimism leaves it up to the people.

Some Questionable Grounds for Pessimism

Not everyone is so hopeful about the utility of trolley cases for resolving the ethical challenges raised by accident scenarios. Trolley Pessimism is the view that there is some mistake in trying to draw on trolley cases to think about the ethics of accident scenarios or the ethics of autonomous vehicles more generally. Different forms of Trolley Pessimism can be distinguished on the basis of what mistake they identify.

One basis for Trolley Pessimism is a distaste for the philosophical method of using thought experiments to arrive at conclusions. Sometimes, this is grounded in the idea that thought experiments that philosophers deploy are so idealized and unrealistic that they are useless when it comes to navigating the real world. We think these sorts of objections rest on a mistaken view of the function and value of thought experiments; we set that aside except to note that a key motivation for Trolley

Optimism is that accident scenarios seem to closely resemble trolley cases. If trolley cases are useless for thinking about accident scenarios it isn't because trolley cases are obviously too unrealistic to be of any use. At the very least, a plausible basis for pessimism must articulate the differences between trolley cases and accident scenarios that prevent us from drawing conclusions about what to do in accident scenarios on the basis of our judgments about trolley cases.

Another basis for Trolley Pessimism tries to do exactly this, to show that there is some point of difference between trolley cases and the behavior of AVs in accident scenarios that makes our verdicts in the former inapplicable to decisions about what to do about the latter. What are the differences between trolley cases and accident scenarios that justify this form of pessimism? Nyholm and Smids (2016) point to several points of disanalogy. For example, in trolley cases, we set aside questions of the moral and legal liability of those that are deciding how to act. The person who will decide whether to divert the trolley, it is assumed, will not be held responsible or liable for whichever choice they make. In the case of accident scenarios, these considerations should inform our deliberations about how AVs should behave in accident scenarios. Another point of disanalogy they raise is that, in trolley cases, the outcomes of various decisions are stipulated to be known with certainty, whereas in the case of accident scenarios, despite what we may want a vehicle to do and given our best efforts, there is some uncertainty about whether its behavior will generate the desired outcome.

Again, we think this is not a plausible basis for Trolley Pessimism. While it is true that traditional trolley cases do stipulate away issues of legal and moral liability and stipulate outcomes with certainty, there is no in principled reason why we can't deploy thought experiments that take these variables into account. We can develop a case that asks what should be done assuming some particular legal liability regime, enumerating the costs to the agent making the decision. Similarly, we can construct a case in which pulling a switch has a 95% chance of altering the course of a trolley and deliberate about whether this alters one's moral obligations. We could even contact the creators of Moral Machine to build these variables into their cases and collect data about what people think should be done in those circumstances and then aggregate that data to dictate the behavior of AVs in accident scenarios.

The Technological Basis for Trolley Pessimism: Lessons from Machine Learning

There is a better basis for pessimism that has its basis in the nature of the enabling technology of autonomous vehicles: machine learning algorithms. For those unfamiliar with machine learning algorithms, we can contrast these algorithms with what we call “traditional algorithms”. An algorithm is simply a set of instructions for executing a task or series of tasks to generate some output given some input. In a traditional algorithm this set of instructions is laid out by hand, each step being explicitly specified by a programmer or designer. In contrast to these traditional algorithms, machine learning algorithms are algorithms that themselves generate algorithms, and these resulting algorithms do not have the steps used to carry out some task specified explicitly by a programmer.

A good analogy for some forms of machine learning, namely *supervised* and *reinforcement* learning, is dog training. Unfortunately, we cannot just program a dog to respond to the words ‘sit’, ‘stand’, ‘stay’, ‘heel’, etc. by wiring up a dog’s brain by hand. Instead, when training a dog, it is common to arrange for situations where the dog will engage in some desired behavior. The dog is then rewarded. For example, a trainer might hold a treat in front of a dog’s nose, lifting it into the air, causing the dog naturally to lift its head and drop its back legs. The dog is then rewarded. After many repetitions, the word ‘sit’ is said right before the treat is lifted. Eventually the dog sits on command, having learned an output for the input ‘sit’.

In the case of machine learning, a programmer can provide a machine learning algorithm with a *training set*, a data set that includes information about which outputs are desirable and which are not. The learner then generates an algorithm that is meant to not only yield appropriate input-output pairs when it is fed inputs that match those in the test set, but to extrapolate beyond the test set, yielding, the programmer hopes, desirable outputs for new input data.

Machine learning is a powerful tool. It allows programmers to develop algorithms to solve problems that would otherwise be extremely tedious or impossible. The AVs likely to be on the road in the foreseeable future will rely on machine learning technologies. At the very least, machine learning is at the heart of the detection systems used in autonomous vehicles. Those detection systems take in input data from various sensors (radar, lidar, cameras) and have to translate that to some output that the car's other systems use to drive the car, to maintain its position within driving lanes, to slow when there is a car in front of it but not when there is merely a line in the pavement.

The fact that AVs depend so heavily on machine learning algorithms grounds a case for Trolley Pessimism. To see why, first take notice of the fact that how an AV behaves in any given accident scenario is mediated by how the algorithm that governs behavior is trained. In order to influence the behavior of an AV in an accident scenario, we will have to do so, in part, by organizing a training set to achieve that behavior. For example, if we want an AV that suddenly confronts the scenario where it must swerve risking harm to its passenger or maintain course and hit some larger number of pedestrians, we might do so by including such scenarios in the training set and marking a particular input-output pair as desirable. This is not the only way we might achieve the desired behavior; the point is simply that behavior in particular scenarios is influenced by choices that programmers and designers make about how to train the machine learning algorithms.

The choices that programmers and designers make about how to train the machine learning algorithms that power AVs involve ethical choices. Programmers will have to make some choices about, for example, what proportion of the training data is dedicated to accident scenarios at all. For example, some programmers might wish to focus on non-accident scenarios or typical driving scenarios, including no data about how a car should behave in accident scenarios. Another might wish to have half the training data dedicated to everyday driving scenarios and half dedicated to accident scenarios. Let's imagine these two programmers are on the same team and arguing about which approach is better, what proportion of the training set should be dedicated to scenarios where the car detects itself to be in an accident scenario where harms can't be avoided. The first programmer argues that the car will very rarely be in those kinds of cases and instead we should train the car for the scenarios it will most likely be in. The second programmer argues that even if the accident scenarios are rare, it's extremely important to make sure the car does the right thing! The first programmer replies that if they dedicate enough of the training set to getting certain behaviors in accident scenarios, it could make the car less safe in typical driving scenarios or even put the car into accident scenarios more often! Clearly this argument over how to train the algorithm that will help govern AV behavior is an ethical argument; it invokes various value-judgments and judgments about how those values are implicated in potential outcomes of the decision to be made.

It follows from the facts that the decisions about how to organize the training regime that yields AV behavior is an ethical decision and that this decision mediates questions about how AVs should behave in particular driving situations, trolley cases do not provide direct guidance about how AVs should behave in accident scenarios, despite any superficial similarities between accident scenarios and trolley cases. There are a several ways to see why.

First, let's suppose that in our imagined argument between the programmers above, we come to believe that the first programmer is correct, that the algorithms that ultimately generate AV behavior should not be generated using any data about accident scenarios. The resulting algorithm will still of course, generate behaviors in such scenarios. The training set just won't have been designed to generate any particular behaviors in those scenarios. In this case, the answer to the question "should we try to model the behaviors of AVs on the verdicts of trolley cases?" is clearly "no!" because we've got good reasons to think we should not try to model the behaviors of AVs in those scenarios at all.

Another way to illustrate the point is to recognize the way trolley cases typically function in ethical theorizing. Trolley cases are thought experiments, imagined examples used to help us test more general principles. Let's imagine we are wondering whether we should accept a principle that we should act in such a way so as to maximize the total number of lives saved (holding fixed things like whether the people whose lives are saved are good people, how large their families are, etc.). Someone asks us to consider the standard trolley case. We imagine a train hurtling down the tracks and must decide whether diverting the trolley onto a track that results in fewer deaths is the right thing to do. Let's assume we come to see this trolley case as lending support to the principle that we really should maximize total lives saved.

If we think that principle is true, it is a principle that programmers and designers should abide by when deciding how to train AVs. The Trolley Optimist might think that the above case justifies us in aiming to ensure that an AV in an accident scenario will not drive into a larger crowd to spare a smaller. However, it could very well turn out that abiding by the principle we've settled on has the implication that we are not justified in doing so. Imagine that in our debate between the programmers above that both are committed to maximizing lives saved. The first programmer argues that they can maximize lives saved by avoiding accident scenarios as much as possible and to do that they should not train the algorithm for accident scenarios at all, but for how to stay out of them. This might have the result that when an AV is in an accident scenario it does veer into a larger crowd to save a smaller, but given that the programmer's decision point is how to program for the whole range of behaviors the car will encounter, they haven't failed to take into account the lesson of the trolley case; they've taken that lesson into account in just the right way. The other programmer might see it is regrettable that the best way to maximize lives saved given the decision they confront will yield this outcome while still acknowledging that this is the approach that conforms with the principle.

To be clear, we are not endorsing any particular view of how AVs should be trained or this particular principle as governing that decision. The point is simply that the Trolley Optimist makes a mistake in thinking that the lesson from trolley cases is a lesson for how an AV should behave in a superficially similar case. The ethical question that designers face is not one about the right thing to do in a specific scenario; it is a question about how to design for the wide-range of scenarios that AVs will find themselves in given that that their choices about how design for one scenario is not isolated from how they choose to design for another.

The upshot of this is not pessimism about the need for ethics in AV design, nor that trolley cases are useless for the task. Instead, the upshot is that we must be much more careful in deploying the resources of ethics, ensuring that we are evaluating the appropriate decision and considering how the technologies at issue relates to the ethical principles and reasoning we hope to deploy. If anything, we hope this paper motivates a closer working relationship between ethicists and designers of AVs to ensure that we are solving the right problems in the right way.

WORKS CITED

Achenbach, Joel (2015). "Driverless Cars are Colliding with the Creepy Trolley Problem," *The Washington Post* December 29, 2015. Accessed November 3, 2018: https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/?utm_term=.8c16a3ee1a28

Doctorow, Cory (2015). "The Problem with Self-Driving Cars: Who Controls the Code?" *The Guardian* December 23, 2015. Accessed November 3, 2018: https://www.theguardian.com/technology/2015/dec/23/the-problem-with-self-driving-cars-who-controls-the-code

Hao, Karen (2018). "Should a Self-Driving Car Kill the Baby or the Grandma? Depends Where You're From," *MIT Technology Review* October 24, 2018. Accessed November 3, 2018: https://www.technologyreview.com/s/612341/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/

Hübner, Dietmar and Lucie White (2018). "Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us Beyond Harm Minimisation," *Ethical Theory and Moral Practice* 21 (3): 685-698.

Lin, Patrick (2013). "The Ethics of Autonomous Cars," *The Atlantic* October 8, 2013. Accessed November 3, 2018: https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/

Marshall, Aarian (2018). "What Can the Trolley Problem Teach Self-Driving Car Engineers?" *Wired* October 24, 2018. Accessed November 3, 2018: https://www.wired.com/story/trolley-problem-teach-self-driving-car-engineers/

Noothigattu, Ritesh, Snehalkumar 'Neil' S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia (2017). “A Voting-Based System for Ethical Decision Making” arXiv.org, submitted September 20, 2017. Accessed November 5, 2018: https://arxiv.org/abs/1709.06692

Nyholm, Sven and Jilles Smids (2016). “The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?” *Ethical Theory and Moral Practice* 19 (5): 1275-1289.

Wallach, Wendell and Colin Allen (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Worstall, Tim (2014). “When Should Your Driverless Car from Google Be Allowed to Kill You?” *Forbes* June 18, 2014. Accessed November 3, 2018: https://www.forbes.com/sites/timworstall/2014/06/18/when-should-your-driverless-car-from-google-be-allowed-to-kill-you/#2e21fba6fa5b