Transcending Trolley Problems:

Reframing and Resolving Ethical Problems for AV/CV Safety

Jason Millar[1] and Ian Kerr[2]

"The way to solve the problem you see in life is to live in a way that makes the problem disappear."

Ludwig Wittgenstein, *Culture and Value*

Policy Provocation

The renaissance of so-called "trolley problems" has been instrumental in bootstrapping discussions on the ethics of automated and connected vehicles (AV/CVs). So much so that AV/CV ethics has become narrowly synonymous with "solving" trolley problems. Policymakers around the world are engaging ethicists to help them solve such problems, the resolutions of which are perceived as necessary preconditions to enabling social trust in AV/CV.

While helpful in ensuring that ethics has a central place in AV/CV policy, the trolley problem's monopoly on AV/CV ethical discourse unhelpfully creates practical engineering and policy challenges. Though trolley problems have provided useful models for evaluating ethical theories and doctrines (e.g., utilitarianism; "the doctrine of double effect"), their practical application to the ethical design of AV/CV has been vastly overstated in some cases and largely misunderstood in others. This mismatch between theoretical and practical relevance has resulted in confusion and frustration among both policymakers and industry players. Responding adequately to the much broader set of issues that comprises AV/CV ethics requires a shift in the way that the policy challenges are framed: rather than framing regulatory responses as trolley problems, policymakers need to mandate transparent processes that generate ethical AV/CV design decisions.

---

[1] Dr. Jason Millar is an Interdisciplinary Postdoctoral Fellow at Stanford University's McCoy Family Center for Ethics in Society, and a Research Affiliate at the Center for Automotive Research at Stanford.

[2] Dr. Ian Kerr is the Canada Research Chair in Ethics, Law and Technology at the University of Ottawa's Faculty of Law.

Objectives

1) To illustrate the philosophical significance of a class of problems referred to as "trolley problems," and assess the merits and demerits of framing problems in AV/CV ethics and policy discussions as trolley problems.
2) To elucidate the practical challenges associated with policies that seek to respond to trolley problems.
3) To sketch an alternative policy option that satisfies the general demand for putting ethical considerations into the design and governance of AV/CVs.

A Very Brief Philosophical History of Trolley Problems

Trolley problems emerged from university philosophy departments in the second half of the 20th century. The original Trolley Problem, in its most popular form[3], asks you to imagine a runaway trolley that is headed toward five unsuspecting people standing on the track. However, it so happens that you are standing next to a switch that, if pulled, would divert the trolley onto a parallel track containing a single unsuspecting person. There is no way to warn the people on either of the tracks. The trolley will undoubtedly kill the five people if left to continue straight, or else it will kill the one person if the switch is pulled. There are no other options. The dilemma entails choosing whether or not to pull the switch and divert the trolley. It asks which of these two awful outcomes we ought to choose.

Thomson's argument was meant to illustrate and probe people's seemingly contradictory ethical intuitions. On the one hand, people seem to agree that one should pull the switch to save the five, but if you ask people whether it is acceptable, for example, to push a large person onto the track to stop the trolley from hitting the five (commonly referred to as the 'Fat Man' version of the trolley problem), they tend to say "no." In the years after Thomson's initial thought experiments, philosophers have made prolific use of trolley problems, posing thousands of variations, all meant to problematize aspects of particular theories, or illustrate interesting intuitive inconsistencies. For Thomson, the purpose behind trolley problems was to solve theoretical problems—in her case, the mental model of the dilemma permitted a moral probe of our intuitions regarding the theoretical distinction between *killing* and *letting die*.

It is critical to note that initial formulations[4] of the trolley problem—as examples in a broader category of philosophical thought experiments—were never meant to answer practical questions about whether one ought to pull a switch, push a button, turn a wheel, or shove a fat man. Nor was the construction of trolley problems ever aimed at shedding light on the design of, or policies governing trolleys. Safety policy for unmanned vehicles had no more to with the problems philosophers were seeking to solve than reducing mortality rates in poisoned cats (Schrödinger's

[3] The original Trolley Problem was created by philosopher Philippa Foot (1967) and was further popularized by Judith Jarvis Thomson's (1976) treatment.

[4] Trolley problems have evolved significantly in the past five decades to the point that some academics speak of "Trolleyology": Altschuler (2014); Edmonds (2014).

Cat[5]), increasing brain longevity (Brain in a Vat[6]), or the permissibility of aftermarket parts for marine vessels (Ship of Theseus[7]). With this we see that the so-called 'Trolley Problem' is in fact a misnomer. It should have been called the 'trolley example.' Neither Foot nor Thomson were seeking to 'solve' anything related to trolleys. They were, in fact, interested in the right and wrong of abortion, the value of utilitarian theory, the doctrine of double effect, and so forth. Trolley scenarios were theoretical heuristics put to theoretical ends. As discussed below, trolley problems were intentionally designed as intractable dilemmas.

The desire to link trolley problems to policy questions about AV/CV came more than 50 years later, when a few well-learned practical ethicists could not resist using them to show that such questions were no longer the luxury of theoretical philosophers. Trolley problems illustrate in the most visceral terms that *ethical design* in the development of AI and robotics has become a modern-day imperative.

How Trolley Problems Became Synonymous with AV/CV Ethics

Today, trolley problems have become synonymous with AV/CV ethics. How can this be, given that they were so clearly directed towards solving theoretical problems?

First, there is relatively direct analogy between runaway trolleys and driverless cars. Trolley problems highlight one category of ethical concerns—safety—which is the primary justification for AV/CVs in the first place (i.e. the promise of fewer road injuries and fatalities). In the early days of AV/CVs, before many were on the road, philosophers and policymakers published a few widely disseminated public thought pieces that made the analogy explicit and argued that AV/CVs turned those academic thought experiments into practical engineering and policy problems.[8]

Second, following the publication of these early articles, and owing to the growing public interest in AV/CVs, there was an explosion of articles in the popular media popularizing and spreading the ethical links between trolley problems and AV/CVs. As any ethics professor will tell you, trolley problems have the ability to capture people's imagination, so the trolley problem took hold in the popular imagination and among engineers and policymakers. A recent Google search on "trolley problem driverless car" returned over 68,000 results.

Finally, near the top of those Google responses, one finds a link to the MIT Moral Machine, a project that has popularized the trolley problem through gamification and claims to offer a solution. In doing so, the Moral Machine acquired a certain infamy among ethicists, engineers

[5] Which was in fact created by Erwin Schrödinger as a counterargument to the "Copenhagen Interpretation" of quantum theory.

[6] Which was in fact created by Hilary Putnam to illustrate Cartesian skepticism.

[7] Which was in fact created by Plutarch to raise questions about the nature or personal identity.

[8] For examples see Goodall 2014; Lin 2013a, 2013b, 2014; Millar 2014a, 2014b.

and policymakers worldwide.[9] The MIT Moral Machine allows anyone to design their own nuanced variation of trolley problem (there are now thousands), and garner public feedback on how people think driverless cars ought to behave in that scenario. The Moral Machine was imagined as a tool to help answer policy questions related to AV/CV ethics. But, for reasons set in the following section, it suffers from major ethical design issues that make any policy or engineering conclusions one might be tempted to derive from it extremely dubious and likely unsound.

Three Practical Problems with Framing AV/CV Ethics as Trolley Problems

The MIT Moral Machine is worthy of further consideration because it helps to underscore ways in which trolley problems create conundrums for AV/CV engineering and policymaking. The MIT Moral Machine takes as its premise the idea that we can design AV/CV ethics and policy (and design ethical automated vehicles) by understanding what the public thinks automated vehicles should do when faced with myriad trolley problems. This might seem like sound evidence-based policymaking and a good basis for AV/CV engineering decision-making—but it is not. Framing AV/CV ethics through this narrow lens engenders three practical problems that generate confusion for engineers and policymakers, struggling to find a way through the ethical murk.

1. Intractability. The whole point of framing trolley problems as ethical dilemmas is to create a (seemingly) intractable problem. In direct juxtaposition to engineering and policy goals, the architecture of trolley problems is intended to ensure that neither of its two lemmas offers a genuinely beneficial outcome. That is, part of the goal of trolley problems is to make them seem irresolvable in any objective sense. From a theoretical perspective, this is a feature of trolley problems—not a bug. It is the intractability of the dilemma that permits theoretical exploration. From the theorist's perspective, this is a good thing since trolley problems are not meant to be solved; quite the opposite, they are merely instruments of thought. It is therefore not surprising that philosophers will not be able to tell you what the right thing to do is in the switch or fat man dilemmas: full stop. Moral disagreement about the outcome is precisely what makes trolley problems such a great philosophical tool. However, it is important to recognize the practical import in this. Framing AV/CV policy options as ethical dilemmas creates problems that don't otherwise exist. Indeed, the best way to escape these intractable scenarios is to engage in ethical design methods that do not give rise to such dilemmas in the first place.

2. Encouraging Argumentum ad populum. The perceived intractability of trolley problems makes it tempting to resolve them by simple appeal to majority opinion. This is as true for trolley problems as it is for any other ethical dilemma, as the MIT Moral Machine so well illustrates. Though policymakers might be tempted to turn to majority opinion for answers in cases where genuine philosophical disagreements exist—as a sort of tie-breaking activity—that approach will offer little in the way of sound justification. Determining whether something is the right thing to do, from the point of view of ethics, is not a question that even the best

[9] See MIT Moral Machine (2016).

polling can answer. Determining whether it is the right thing to do to requires, most importantly, *sound ethical deliberation and argumentation*. It might be the case that the public agrees with the outcome of sound ethical deliberation, though in many cases they will not. The point is that ethical justifications rest on good arguments. Public polling does not. Determining what is the right thing to do is simply not a matter of opinion polling.

3. Frustrating Important Practical Decision-making. Finally, treating AV/CV ethics as synonymous with trolley problems risks alienating engineers and policymakers who would otherwise be inclined to engage in meaningful ethical design. Trolley problems represent extreme engineering corner cases of a sort are simply not a priority for today's AV/CV designers and manufacturers. Constantly distracting engineers and policymakers with trolley problems—especially in light of their intractability and appeal to populism—can have a serious frustrating effect on actual ethical design objectives. AV/CV ethics encompasses a set of problems much broader than those illustrated by way of trolley problems. The list of serious and significant ethical issues that designers need to face in the near term includes:

- how or whether to 'handover' the driving function in certain emergency situations;
- how and when to ensure transparency in design when the design significantly impacts the public good;
- how to ensure manufacturers are striking the right balance between private and public interests in redesigning the mobility system;
- whether or when AV/CVs should be permitted to operate without safety drivers or passengers;
- whether there should be default ethical settings subject to change by consumers and, if so, what those defaults should be;
- how or whether to replicate (or replace) informal interactional norms that currently exist between pedestrians and vehicles;
- how to ensure privacy in the age of connected vehicles;
- how to structure a mobility system that situates an unprecedented level of control in the hands of software designers and engineers; and
- whether or how to ensure that the emerging AV/CV mobility system respects democratic norms such as, provincial or municipal mobility (and city) planning objectives, individual autonomy and freedom, and so on.[10]

These and many other pressing and substantial ethical issues implicate AV/CV safety and are much more tractable than trolley problems. Asking engineers and policymakers to prioritize and solve no-win speculative scenarios risks inattentiveness to on-the-ground ethical issues that require immediate consideration.

[10] Millar 2017.

The Trolley Problem's Early Impact on AV/CV Policy

Despite frustrations, and despite the philosophical and practical challenges that trolley problems pose for engineers and policymakers, there have been at least two major and direct international policy responses to trolley problems to date: the US Department of Transportation (DOT) and National Highway Traffic Safety Administration's (NHTSA's) AV/CV Policy documents[11] and Germany's legislative AV/CV initiatives[12]. These responses were largely motivated by the popularization of trolley problems as being synonymous with AV/CV ethics.

In 2016, the US DOT and NHTSA released their first AV/CV policy. In a nutshell, the policy required AV/CV manufacturers to demonstrate that they had addressed fifteen areas in the design and manufacture of any automated driving system. One of those areas was "ethical considerations," and the policy made it clear in the preamble on ethical considerations that the ethical concerns NHTSA had were a direct response to growing concerns over trolley problems. However, NHTSA stopped short of explicitly addressing any particular ethical issue, instead leaving it to industry to provide solutions to whatever ethical challenges were implicated by their new AV/CV technologies.

In 2017, the policy was updated and the section requiring manufacturers to demonstrate "ethical considerations" removed. However, the Frequently Asked Question section of the revised NHTSA AV/CV policy clearly specifies:

> "Ethical considerations are essential to automated driving technology development. However, currently, there is no consensus around acceptable ethical decision-making given the depth of the element is not yet understood nor are there metrics to evaluate against. NHTSA plans to work with industry, States, and safety advocates to further research the establishment of an industry developed framework for addressing ethical considerations and fostering transparency in automated driving technology decision making. The Agency will also collaborate with industry to develop standard test and simulation scenarios that culminate in an ethical decision."[13]

Thus, it is clear that the "ethical considerations" requirement was not removed for lack of relevance to NHTSA's mandate. Rather, the removal was a response to uncertainty of what, precisely, ethical considerations amount to in the design, manufacture and governance of AV/CV technology. There is little doubt that the problems discussed above directly contribute to this uncertainty.

The German case is different. In 2017, resulting from the work of a federally appointed national ethics committee, Germany released the first code of ethics for AV/CV. Unlike the US approach,

---

[11] NHTSA 2017a, 2016.

[12] Luetge 2017.

[13] NHTSA 2017b.

Germany opted to address features common to trolley problems head-on. The German approach sets out several overarching ethical principles intended to eliminate certain design options that might otherwise seem ethically justifiable. Chief among them are the following[14]:

1. "Technological development obeys the principle of personal autonomy, which means that individuals enjoy freedom of action for which they themselves are responsible.
2. Within the constraints of what is technologically feasible, the systems must be programmed to accept damage to animals or property in a conflict if this means that personal injury can be prevented.
3. In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical, or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable. Those parties involved in the generation of mobility risks must not sacrifice non-involved parties."

These principles directly address features specific to trolley problems in the way that they refer to trade-offs between persons and property, persons and non-persons, and different kinds of persons. As aspirational or hortatory principles, they are certainly laudable; they importantly extend to the AV/CV setting values that we would not wish to lose in the delegation of the driving task to machines. But it is unclear from an engineering or policy-making perspective whether they could ever in fact be achieved. It is perhaps equally unclear whether the human tendency to 'swerve'—which is if course the *fundamental presumption* in the trolley car analogy—will be equally salient in level 5 autonomy crashes in the future. For all we know, the better ethical approach, from an engineering safety perspective, will be to slow driving speeds, massively increase the effectiveness of sensors and actuators in anticipating and responding to the need to brake, and dedicate different lanes for different roadway uses. The future of swerving—like the trolley problem itself—remains uncertain.

One thing is clear. The contrast between the US and German approaches reflects obvious differences in attitudes toward regulation in general. The US approach leaves almost all ethical decision making to industry, whereas the German approach dives head-first, unafraid to adjudicates the specifics of trolley problems. Considering which style of approach Canada should adopt is itself is a more tractable and worthwhile ethical question than trolley problems themselves. From a Canadian perspective, the current US approach has too little to say on ethics, the German approach, perhaps, too much.

Transcending Trolley Problems – Process-Focused Policy

A growing number of researchers, the authors of this policy provocation among them, are actively investigating ethical frameworks without paying much head to trolley problems. The

[14] Luetge 2017.

policy challenge is to find ways of practically addressing as broad a range of ethical issues raised by AV/CVs as possible.

Although we have enumerated a number of pressing and substantial ethical issues that take precedence over trolley problems, the purpose of this policy provocation is not to discuss any of them on a substantive level. Having argued against narrowly framing AV/CV ethics as a need to respond directly to trolley problems, we end by encouraging a broad, process-focused approach.

For current purposes, this approach is agnostic as to the particular ethical issue being addressed, or process being implemented. It focuses instead on the concept of *ethical due diligence*. Ethical due diligence is an approach that is used in other domains involving difficult practical ethical challenges, such as research and healthcare. We believe that a good first step in the creation of ethical AV/CV is to mandate an ethics-oriented process in the design of AV/CVs.

According to this approach, *AV/CV manufacturers would be required to demonstrate that they have put in place meaningful ethical design policies and practices in order to identify and address, with reasonable consideration and justification, ethical issues implicated by the particular technologies they are producing*.

This policy approach encourages and allows manufacturers to develop robust and ethically-oriented processes in attempts to respond to substantive ethical issues. Some examples: a manufacturer could adopt ethical design thinking; implement internal ethics reviews as part of the design process; engage ethicists to help identify and think through ethical challenges posed by their technology, and so on. This policy approach encourages a variety of ethical solutions, the governance focus shifting to requirements around the processes as a means of demonstrating that reasonable ethical considerations and justifications are produced for those design features that have considerable ethical implications for users and society.

As an example, the following policy proposal could serve as the *starting point for a reasonable discussion of how best to approach the development of ethical AV/CV.*

When constructing policy for AV/CV ethics, consider the following guiding principles:

(1) Focus on mandating ethics-oriented processes that require industry to identify and attend to emerging ethical issues in AV/CV such that design decisions show evidence of reasonable ethical consideration and justification.

(2) Ensure a reasonable level of transparency such that a manufacturer's ethics-oriented processes, and the decisions they produce, can be scrutinized by relevant regulators and the public where necessary.

These guiding principles are meant as the first—and not the last—words on AV/CV ethics. Regulators might one day determine that ethical settings are required and particular defaults established. Other ethical design features might also be mandated to promote democratic values, safety, or the public good. But these starting points are sure to take us further down the

road to ethical AV/CV than remaining stuck at the switch on the trolley tracks. Trolley problems jumpstarted the ethical discussion around AV/CV. Now it is time to transcend them.

References

Altschuler, G.C. (2014). "Trolleyology" *Psychology Today* Online: https://www.psychologytoday.com/blog/is-america/201401/trolleyology

Edmonds, D. (2014). *Would You Kill The Fat Man? The Trolley Problem and What Your Answer* Tells *Us About Right and Wrong* (Princeton: Princeton University Press).

Foot, Philippa. (1967). "The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices." *The Oxford Review* 5:5-15.

Goodall, Noah J. 2014. "Ethical Decision Making During Automated Vehicle Crashes." *Transportation Research Record: Journal of the Transportation Research Board* 2424: 58-65.

Lin, P. (2013a). "The Ethics of Saving Lives with Autonomous Cars Is Far Murkier Than You Think." *WIRED* (Jul 30). Online: https://www.wired.com/2013/07/the-surprising-ethics-of-robot-cars/.

--. (2013b). "The Ethics of Autonomous Cars." *The Atlantic* (Oct 8). Online: http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/.

--. (2014). "The Robot Car of Tomorrow May Just Be Programmed to Hit You." *WIRED* (May 6). Online: http://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/.

Luetge, C. (2017). "The German Ethics Code for Automated and Connected Driving." *Philosophy & Technology* 30:547-558.

Millar, J. (2014a). "Should your robot driver kill you to save a child's life?", *The Conversation UK* (Aug 1). https://theconversation.com/should-your-robot-driver-kill-you-to-save-a-childs-life-29926.

--. (2014b). "You Should Have a Say in Your Robot Car's Code of Ethics." *WIRED* (Sept 2). Online: http://www.wired.com/2014/09/set-the-ethics-robot-car/.

--. Millar, J. (2017). "Ethics Setting for Autonomous Vehicles." In P. Lin, R. Jenkins & K. Abney (Eds.). *Robot Ethics 2.0 From Autonomous Cars to Artificial Intelligence.* (Oxford: Oxford University Press).

MIT Moral Machine. (2016). http://moralmachine.mit.edu.

NHTSA. (2016). *Federal Automated Vehicle Policy: Accelerating the Next Revolution in Roadway Safety*. Online: https://www.transportation.gov/sites/dot.gov/files/docs/AV%20policy%20guidance%20PDF.pdf

--. (2017a). *Automated Driving Systems 2.0: A Vision for Safety*. Online: https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf.

--. (2017b). "Automated Driving Systems: FAQ." *NHSTA.* Online: https://www.nhtsa.gov/manufacturers/automated-driving-systems#automated-driving-systems-faq.

Thomson, Judith Jarvis. (1976). "Killing, Letting Die, and the Trolley Problem." *The Monist* 59:204-217.