Ethics of Autonomous Vehicles

Introduction

Every year, 1.35 million people die from road crashes worldwide. Self-driving cars, or autonomous vehicles (AVs), will not only reduce this rate drastically, but they will also fundamentally transform how people live and work. They will shorten commute times, enable people to live further from cities, reduce harmful gas emissions, and free up road and parking space for other uses. As implement this technology, we will confront a myriad of ethical challenges.

Why Trolley Problems are Morally Relevant

There is a view that trolley problems – scenarios where a vehicle must "choose" one of several options that will cause harm in a collision – are not relevant for the ethics of self-driving cars. This view is typically defended through one of four arguments. Below I will summarize each of these arguments, as well as University of Cambridge researcher Geoff Keeling's counterarguments.

Argument 1: "Not Going to Happen Argument"

Self-driving cars will not encounter trolley problem-like cases. Therefore, trolley problems are not relevant for the development of the ethics of self-driving cars.

Counterarguments to Argument 1

Counterargument 1

It is possible that AVs will encounter trolley problem-like scenarios, especially if the use of the vehicles becomes widespread. For example:

> *"The AV is travelling on a two-lane bridge. A bus in the other lane swerves into the AV's lane. The AV can either brake, in which case it will collide with the bus; or it can swerve into the other lane, in which case it will hit the side of the bridge. Cases like these seem plausible" (Keeling).*

Additionally, "absence of evidence is not evidence of absence." One cannot rule out the possibility of an AV encountering a trolley problem-like scenario, even if it has not occurred in the past.

Counterargument 2

Even if AVs do not encounter trolley problem-like scenarios, such cases may still be relevant. Studying idealized scenarios can provide valuable insight, even if the real world presents non-idealized scenarios. For example, physicists and chemists use the ideal gas law (PV=nRT) to study the fundamental nature of gases, even though gases exhibit non-ideal behaviors in practice. In the context of self-driving cars, studying idealized trolley problems can provide insight into topics such as whether there is a moral difference between killing and letting die. Such insight can be applicable towards the ethics of self-driving cars, even in cases that don't resemble an idealized trolley problem.

For example, non-literal trolley problems could arise in the form of balancing the safety of walkers and bikers with the interest of the passenger. How fast should an AV drive with pedestrians nearby? As ethics and policy professor Johannes Himmelreich described:

"As soon as a car goes faster than walking pace, it is unable to prevent from crashing into a child that might run onto the road in the last second. But walking pace is, of course, way too slow. Everyone needs to get to places. So how should engineers strike the balance between safety and mobility? And what speed is safe enough?" (Himmelreich).

The Economist also provided an example in a 2018 article:

"If you stay relatively near to the cycle lane, you're increasing the chance of hitting a cyclist, but reducing the chance of hitting another car in the next lane over...Repeat that over hundreds of millions of trips, and you're going to see a skew in the [accident] statistics." (Economist).

Trolley problem ethics could shed light on ways to resolve these ethical tradeoffs.

Argument 2: "Moral Difference Argument"

There is a moral difference between trolley cases and the real-world cases a self-driving car would face.

In trolley problems, there is no information on who is guilty, and none of the actors have moral obligations. In contrast, with self-driving cars, one can examine the events leading up to the case, and thus determine who is responsible. Additionally, the manufacturers might have legal obligations for the welfare of the AVs' passengers. These factors, which are not present in trolley problems, influence the moral permissibility of the AVs' behaviors.

Another difference between trolley problems and the situations self-driving cars would face is probability. Trolley problems depend on certainty of outcome. But with self-driving cars, the consequences of each course of action can only be estimated with a probability.

Counterargument to Argument 2

For the guilt-responsibility difference between trolley problems and cases AVs would encounter, this is merely an additional moral property present for AVs that is not present in trolley problems. An extra moral property does not diminish the relevance of the other moral properties shared by both trolley problems and self-driving cars. The moral properties that are shared can be studied, irrespective of the moral properties that are not shared.

For the certainty of outcome difference between trolley problems and the cases AVs would encounter, this changes the moral calculation process. Yet, that does not render trolley problem cases irrelevant. With a scenario that involves risk and probability, the self-driving car can take both the probability and severity of each outcome into account when "calculating" which course of action to take. This is an added layer of complexity, yet the general moral questions are the same. Thus, one can still analyze trolley problem scenarios to address the moral questions of self-driving cars.

Argument 3: "Impossible Deliberation Argument"

Self-driving cars cannot solve moral dilemmas based on trolley problem cases because they typically use a bottom-up approach to decision-making, whereas trolley problems are based on a top-down approach.

In a top-down approach (as in trolley problems), the course of action is determined by explicit rules about the outcome.

In a bottom-up approach, the AV's course of action is guided by connectionist learning algorithms – convolutional neural networks that mimic the human brain. The AV learns and optimizes its behavior continuously. Ultimately, its decision-making process is abstract and non-linear. Since the self-driving car programmers will not directly code the responses to each possible scenario, analyzing trolley problem cases isn't practically useful.

Counterargument to Argument 3

When designing self-driving cars, developers can use a "value-sensitive design process." Engineers, stakeholders, regulators, and moral philosophers would work cooperatively to determine the ethical implications of the design choices for AVs. These technological specifications would be applied to ensure the "final product" reflects our moral values. Discussions such as the trolley problem can be used to help designers better understand the ethics behind the decisions AVs will need to make.

For example, discussions on the trolley problem led Frances Kamm to develop the Principle of Permissible Harm. This principle describes patterns in our intuitions in trolley cases, which provides insight into what deems an act morally permissible or impermissible. Thus, the trolley problem is relevant for the moral theories that emerge, since those ideas can be applied towards the value-sensitive design process of self-driving cars.

Argument 4: "Wrong Question Argument"

"The values which ought to be encoded into AV decision-making algorithms are not determined by moral considerations." Trolley problems may provide moral insight into AV collisions, but that isn't what we need.

For any trolley problem solution, many people in society will disagree with underlying moral principles. The issue is achieving broad societal acceptance for the solution. For self-driving cars, people will not comply with a system where they disagree with the moral values encoded. As long as this barrier exists, the trolley problem is irrelevant.

Counterargument to Argument 4

In general, people usually accept legal moralism, in which laws (such as that it is illegal to steal) are based on moral grounds. Because the ethics of self-driving cars are nuanced, there will be people who disagree with their legal-ethical framework. However, it will not be at such a level that broad social acceptance would become an issue.

Additionally, although broad social acceptance is a factor in the AV design process, it is not *primarily* a social dilemma. If public opinion holds that AVs should behave according to immoral principles, broad social acceptance is insufficient. But if the public's view of how AVs should behave properly aligns with good moral principles, AVs' choices should reflect those views. Both of these conclusions are based on the moral principles, not their social reception. Thus, the social acceptance factor is a distinct issue from the moral considerations of trolley problem cases. Although trolley problems cannot solve *every* aspect of the development of self-driving cars (such as the social ones), they can provide insight into the moral challenges.

Works Cited

Himmelreich, Johannes. “The Everyday Ethical Challenges of Self-Driving Cars.” *The Conversation*, 27 May 2018, theconversation.com/the-everyday-ethical-challenges-of-self-driving-cars-92710.

Keeling, Geoff. “Why Trolley Problems Matter for the Ethics of Automated Vehicles.” *Science and Engineering Ethics*, Springer Netherlands, 4 Mar. 2019, link.springer.com/article/10.1007/s11948-019-00096-1.

“Whom Should Self-Driving Cars Protect in an Accident?” *The Economist*, The Economist Newspaper, 27 Oct. 2018, economist.com/science-and-technology/2018/10/27/whom-should-self-driving-cars-protect-in-an-accident.

Deontological Approach

In a deontological approach, self-driving cars would follow duty-bound principles, regardless of the consequences, to determine the most moral course of action in trolley problem-like scenarios. To analyze this, I will focus on one simple trolley problem scenario – the image displayed in the middle of the poster. Assuming the self-driving car cannot brake, should it continue on its natural course and let five people die? Or should it intervene to change its course and kill one person?

According to the most basic form of deontological ethics, the car would continue on its natural, default course and let five people die. The AV must follow duty-bound principles (perhaps defined by Kant's categorical imperative) such as "do not kill."

In a more sophisticated form of the deontological argument, one could claim that there is an intrinsic difference between killing versus letting die. Killing is a violation of a negative duty to refrain from killing. Letting die is a violation of the positive duty to save lives. According to philosopher Raymond A. Belliotti, the following are the possible relationships between positive and negative duties:

1) "All negative duties are equally obligatory, and to violate a negative duty is a morally worse act than to violate a positive duty."
2) "Any violation of a negative duty is a morally worse act than any violation of a positive duty."
3) "Any violation of a particular negative duty is a morally worse act than any violation of its correlated positive duty."
4) "Some violations of a particular negative duty are morally worse acts, *ceteris paribus*, than some violations of its correlated positive duty."
5) "Any violation of a particular negative duty is a morally worse act, *ceteris paribus*, than any violation of its correlated positive duty" (Belliotti).

If violating the negative duty to not kill is an intrinsically worse act than the violating positive duty to save lives, then the self-driving car should let five people die rather than kill one person.

As one takes the deontology approach to the limits, it becomes more difficult to fully embrace deontology. One might ask the following:

- Can the self-driving car switch its course to save 10 lives? 100? 1,000?

- Can a self-driving car break the law (e.g. cross a double yellow line, run a red light, etc.) to save lives or mitigate the risk of killing?

A strict deontologist would answer “no” to those questions, unless they had other Kantian duties that outweighed the duty to not kill.

Subjectivity is another factor that makes it difficult to fully embrace deontological ethics. Who gets to decide what duty-bound principles are “correct”? Even if people agree upon the letter of the moral principles, there may be disagreement on what those principles entail. In this example, there may be disagreement on whether “do not kill” encompasses preventing death from occuring.

Works Cited

Raymond A. “Killing, Letting Die, and Thomson.” *Crítica: Revista Hispanoamericana De Filosofía*, vol. 14, no. 40, 1982, pp. 61–74. *JSTOR*, jstor.org/stable/40104267. Accessed 26 May 2020.

Cooper Hewitt, Smithsonian Design Museum. “Moral Machine, 2016.” *Cooper Hewitt Collection*, Smithsonian Institution, 2016, images.collection.cooperhewitt.org/344850_174e2aab066767e4_b.jpg.

Utilitarian Approach

In a utilitarian approach to trolley problem cases, self-driving cars would determine the most moral course of action based on what allows "the greatest amount of good for the greatest number of people." To analyze this, I will focus again on the simple trolley problem scenario – the image displayed in the middle of the poster. Should the self-driving car continue on its natural course and let five people die, or should it intervene to change its course and kill one person?

According to the most basic form of utilitarian ethics – Bentham's act utilitarianism – the car should choose the action that minimizes total harm and produces the best possible outcome. It should change its course and kill one person to save the other five. Here, unlike in the deontological approach, action is morally equivalent to inaction, as they have the same consequence.

Another form of utilitarianism is rule utilitarianism, where morality is based on choosing the action that aligns with the rule that produces the best possible outcome. This theory overlaps with deontology, both being rule-based. However, deontological duties stem from something intrinsic and absolute, while rule utilitarianism is based on rules that have utility.

In "The Moral Landscape," neuroscientist and philosopher Sam Harris offers a more sophisticated form of utilitarian ethics. Harris argues that morality is based on consciousness. In a "space of peaks and valleys, where the peaks correspond to the heights of flourishing possible for any conscious system, and the valleys correspond to the deepest depths of misery," the goal of morality is to maximize the peaks and minimize the valleys (Harris). The answers to ethical questions should be whatever realizes that goal. If humans can measure and quantify people's conscious states – which is becoming possible through technologies such as fMRI – theoretically, we can apply science (genetics, neuroscience, economics, sociology, etc.) to answer ethical questions.

Harris is arguing that the gap between philosophy and science is the ability to collect empirical data. With the ability to objectively understand states of consciousness (epistemology), moral philosophy is becoming a science. This aligns with the theory of positivism – that knowledge is based on natural phenomena and studied through science and logic.

Harris's theory relates to the utilitarian approach to solving AV trolley problem scenarios. It calls for a broader account of the consequences of each course of action, beyond just the

number of deaths. It accounts for every experience by conscious beings that was affected by the course of action. This includes the emotional trauma inflicted on the bystanders, the car passenger, and the passenger's family members. It includes the legal issues caused by the collision. It includes the ramifications of making exceptions to principles such as natural rights and equal protection (which inevitably occurs when it is permitted to kill one innocent person to save five). It includes the differences in how the death of one person would impact aggregate wellbeing versus the death of another person. For example, the death of an important political or business figure might be worse than the death of a homeless person. The death of a person with many loved ones might be worse than the death of a person with few loved ones. Yet, the act of assigning people moral values, which determine the importance of preserving them in trolley cases, might bring negative ramifications (e.g. privacy issues, emotional discomfort) that outweigh the first-order benefits.

Physics professor Sean Carrol highlighted several issues with Harris's Moral Landscape. These issues have implications for the Moral Landscape's ability to determine self-driving car ethics. His central argument was that "you can't derive ought from is" (Carrol). Performing science, such as analyzing people's conscious experiences, is observation of empirical reality. In contrast, morality is describing how society should be set up to create conditions that maximize well-being. This idea relates to Hume's idea of the is–ought problem: one cannot easily move from descriptive statements to prescriptive statements. Morality and science remain separated by the is–ought gap.

Carrol further clarified his argument, stating three reasons for why the is–ought gap exists in this instance. The first is that "there's no single definition of well-being" (Carrol). One cannot perform a scientific experiment that verifies what the correct definition is. Harris believes he has the definition, but others might dispute that.

Second, "it's not self-evident that maximizing well-being, however defined, is the proper goal of morality" (Carrol). Harris's definition can be justified philosophically and rationally, but not empirically. Philosophers such as Dostoevsky have argued that a utopia, where well-being and rationality are maximized, is undesirable. Additionally, nobody has *empirically* proven that utilitarianism is more correct than deontology, and deontology isn't concerned with maximizing well-being.

Lastly, "there's no simple way to aggregate well-being over different individuals" (Carrol). Different people's interests often conflict with one another. This is true not only between

different humans, but also between humans and other animal species. Solving these conflicts and calculating the moral correctness of different courses of action is impossible both in practice and in principle, Carrol claims.

Carrol's counterarguments to Harris's Moral Landscape appeal to moral relativism – that moral judgements are subjective, not universal. This disagreement highlights the philosophical debate between relativism and absolutism.

Works Cited

Carroll, Sean. “Science And Morality: You Can't Derive 'Ought' From 'Is'.” *National Public Radio*, 4 May 2010, npr.org/sections/13.7/2010/05/04/126504492/you-can-t-derive-ought-from-is.

Cooper Hewitt, Smithsonian Design Museum. “Moral Machine, 2016.” *Cooper Hewitt Collection*, Smithsonian Institution, 2016, images.collection.cooperhewitt.org/344850_174e2aab066767e4_b.jpg.

Harris, Sam. “A New Science of Morality, Part 3.” *Edge.org*, Edge Foundation, Inc., 2010, edge.org/conversation/sam_harris-a-new-science-of-morality-part-3.

Harris, Sam. “Clarifying the Moral Landscape.” *Sam Harris*, 4 Dec. 2017, samharris.org/clarifying-the-landscape/.

“Random” Approach

In a random approach, self-driving cars would determine the most moral course of action based on random selection. This approach, initially proposed by Kyoto University informatics researchers Liang Zhao and Wenlong Li, is far less mainstream than the deontological and utilitarian approaches.

I will focus again on the simple trolley problem scenario – the image displayed in the middle of the poster. Should the self-driving car continue on its natural course and let five people die, or should it intervene to change its course and kill one person?

In the random selecting approach, the car would randomly select between continuing on its natural course, thereby letting five people die, and intervening to change its course and kill one person.

Only after considering the shortcomings of the other two approaches does random selecting appear reasonable. In the deontological approach, it is naive to allow harm to occur to five people based on the negative duty to not kill. By “doing nothing” (allowing the car to continue on its default path), a choice is still being made. The car is still killing people, even if it wasn’t changing its path. Another general shortcoming of the deontological approach, as Zhao and Li argue, is that there is no duty-bound principle that works for all situations. There will always be exceptions to the rules, as well as disagreement on what the rules should be.

The utilitarian approach also has its drawbacks. First, one cannot ever establish scientific certainty about which course of action maximizes utility. The calculation of each course of action’s utility is too complex, involving many qualitative factors. Second, the utilitarian approach’s ranking system (prioritizing what types of people should be saved in collisions) poses other ethical challenges. One one hand, young and healthy people should be preserved over the old and sick. People with important business or political roles should be prioritized, since other people depend on them. But as Zhao and Li describe, “this approach leads to challenges to other more fundamental principles such as ‘Everyone should be equal’” (Zhao and Li 3). The families and friends of collision victims would know that their loved one was killed by a computer algorithm deliberately designed by someone else. People would feel discriminated against, subjected to a system that systematically favors or disfavors them.

The most significant benefit for the random approach is that it is practical. After the first few years of implementation, society can reevaluate whether the utilitarian or deontological approach is better based on experience. During that time, we would also learn how rare trolley

cases are, which would impact how comfortable we are accepting random selecting as a permanent solution.

Another benefit of the random approach is that nobody is to blame for collisions. People wouldn't be subjected to an algorithm that favors or disfavors them. Zhao and Li argue that "random choice is similar to the case of negligent homicide, whose definition is the killing of another person through gross negligence or without malice… It is quite different to kill someone because of bias, which may equal to murder" (Zhao and Li 4).

I found one issue with random selecting that Zhao and Li failed to address. Sartre, a key figure in existentialist philosophy, argued, "if I do not choose, I am still choosing" (Sartre 54). One cannot opt out of moral choice, as that is a choice in itself. The random selecting approach is more disconnected from direct choice than the deontological approach, but it still runs into the same inescapable problem of choice.

Works Cited

Cooper Hewitt, Smithsonian Design Museum. “Moral Machine, 2016.” *Cooper Hewitt Collection*, Smithsonian Institution, 2016, images.collection.cooperhewitt.org/344850_174e2aab066767e4_b.jpg.

“The Humanism of Existentialism.” *Essays in Existentialism*, by Jean-Paul Sartre, Citadel Press, 1965, p. 54.

Zhao, Liang, and Wenlong Li. “’Choose for no choose’ — Random-Selecting Option for the Trolley Problem in Autonomous Driving.” *Ethics in AI, College Park, Maryland, USA, July 2019,* Kyoto University, July 2019.

Who gets to Decide the Ethical Framework?

Unlike normal cars, self-driving cars enable the car designers and manufacturers to determine the ethics of the vehicles. This raises the questions – how much control should the technology corporations have, and who should have a voice in the moral design process?

Inevitably, the technologists and designers will have significant control. But ideally, based on Rousseau's idea of democracy – that law should be based on the general will of the society – the public should also have some input.

Another question that would influence the development of an ethical framework of AVs is of the nature of moral truth. One view is ethical absolutism – that there are universally valid and invalid truths about what is right and wrong. According to this view, there should be a single set of ethical rules for all self-driving cars. This makes it difficult to rely on surveys, because the responses to ethical questions would vary between different societies. Thus, an absolutist might rely more heavily on the input of experts (technologists, philosophers, scientists) than the public's general will.

According to ethical relativism, moral truths are only true relative to a society. The ethical norms of one society could be different than those in another society, yet both could be valid. This theory would permit different companies and countries to decide what ethics are encoded in their self-driving cars. An ethical relativist could value public opinion more heavily, since they accept that ethical norms can vary by society while still being valid.

Although absolutism can be connected to relying on experts and relativism can be connected to relying on the public, the connection isn't perfect. In other words, even if someone confirmed whether ethical absolutism or relativism is true, that doesn't necessarily mean we should rely solely on experts or the public, respectively.

For ethical absolutism, different experts may have contrasting views on the same ethical issue. Even if agreement is met between all people, that agreement might be different from the ethical views that succeed it in the future. Also, people's views of the truth will only ever be epistemology. We can never confirm with certainty whether our epistemological understanding aligns with the "absolute" ontology. These nuances make it difficult to develop an ethical framework under absolutism.

For ethical relativism, the public may hold views that are truly harmful in practice. For example, people might believe that slavery is a positive institution, or that we should hunt for witches. Relativism does not allow for the refutal of these dangerous ideas. Additionally,

relativism can be self-contradicting. If no view can be definitively rejected, a relativist cannot rule out the possibility that absolutism is correct.

One theory that combines absolutism and relativism is ethical pluralism. Pluralism holds that there are absolute truths, yet those truths can be understood and actualized in different ways. There can also be multiple truths that conflict with one another but are simultaneously correct. According to this view, there is not a single valid approach for designing the ethical framework of self-driving cars, but there can be invalid ones.

Work Cited

Karnouskos, Stamatis. “Self-Driving Car Acceptance and the Role of Ethics”. *IEEE Transactions on Engineering Management,* 2018, *ResearchGate,* DOI: 10.1109/TEM.2018.2877307.

MIT Moral Machine

MIT's Media Lab conducted a survey experiment called "Moral Machine" to provide insight into people's ethical priorities for self-driving cars. Using a game-like interface, the experiment showed people trolley problem-like scenarios and asked them how the self-driving car would respond.

In the results, there was general agreement on just these two principles: prioritize saving more lives over less lives, and prioritize saving humans over animals. For everything else, there was strong disagreement between different countries. The following are some of the points of disagreement:

- Prioritize saving passengers or pedestrians?
- Prioritize saving younger people or older people?
- Prioritize saving women or men?
- Prioritize saving healthy or sick people?
- Prioritize saving higher social status or lower?
- Prioritize saving law-breakers or law-abiders?
- Should the car change its course (take action) or stay on course (inaction)?

The image displayed in the bottom right of the poster illustrates these preferences by region.

The results of the Moral Machine experiment highlight how complex the moral design process is. Obviously, there should also be reasonable boundaries, determined by both the general will of the public (as determined by surveys) and experts (moral philosophers, stakeholders, technologists, etc.). Yet because the preferences in the survey were so mixed, I believe individuals should be able to customize certain settings. The ability to make decisions within ethical certain constraints is not much of a change from the present driving situation. People are already required to operate within the constraints of the law while maintaining responsibility over their own ethical choices. My solution would preserve this precedent.

Works Cited

Cooper Hewitt, Smithsonian Design Museum. “Moral Machine, 2016.” *Cooper Hewitt Collection*, Smithsonian Institution, 2016, images.collection.cooperhewitt.org/344850_174e2aab066767e4_b.jpg.

Hao, Karen. “Should a Self-Driving Car Kill the Baby or the Grandma? Depends on Where You're from.” *MIT Technology Review*, 24 Apr. 2018, technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem.

Intuition vs. Conscious Reasoning

One of the foundational philosophical debates is how truth is determined. Philosophers such as David Hume and John Locke advocated for empiricism – gathering truth from experience and the senses. Philosophers such as Descartes and Plato believed in rationalism – that truth should be determined by logic and reasoning.

In developing the ethical framework of self-driving cars, society must determine whether to rely on empirical knowledge (our moral intuitions and instincts) or conscious reasoning. Ideally, our intuitions and conscious reasoning would agree. But they don't always.

Consider the following two versions of the trolley problem (not for self-driving cars, but for actual trolleys), discussed by Peter Singer:

> *"In the standard trolley problem, you are standing by a railroad track when you notice that a trolley, with no one aboard, is rolling down the track, heading for a group of five people. They will all be killed if the trolley continues on its present track. The only thing you can do to prevent these five deaths is to throw a switch that will divert the trolley onto a side track, where it will kill only one person. When asked what you should do in these circumstances, most people say that you should divert the trolley onto the side track, thus saving four lives.*
>
> *In another version of the problem, the trolley, as before, is about to kill five people. This time, however, you are not standing near the track, but on a footbridge above the track. You cannot divert the trolley. You consider jumping off the bridge, in front of the trolley, thus sacrificing yourself to save the imperiled people, but you realize that you are far too light to stop the trolley. Standing next to you, however, is a very large stranger. The only way you can stop the trolley killing five people is by pushing this large stranger off the footbridge, in front of the trolley. If you push the stranger off, he will be killed, but you will save the other five. When asked what you should do in these circumstances, most people say that you should not push the stranger off the bridge"* (Singer 339-340).

When asked, people cannot explain why they chose to divert the trolley in the first version but wouldn't push the stranger off the bridge in the second version. Although this example is with trolleys rather than self-driving cars, the underlying problem is still relevant: our intuitions don't always align with the principles we formulate through conscious reasoning.

Therefore, the question of whether to rely on intuition or conscious reasoning is relevant to the moral design process of self-driving cars.

Modern psychologists have explored deeply the tension between intuition and conscious reasoning in ethics. Jonathan Haidt, who has done experimental research on the subject, said "moral reasoning is generally done post-hoc, to search for confirmation of our fast, automatic intuitive responses. I am therefore skeptical of the power of reasoning to bring us to the right conclusions" (Haidt). Based on Haidt's argument, we must rely on intuition to a certain extent, because reasoning is merely a reflection of our intuition.

Psychologist Joshua Greene has taken the opposite stance. He argued that our moral intuitions are the result of millions of years of evolution. Although that evolution has provided us with good intuitions, it has also led to tribal instincts. Our moral instincts are the driving factor in atrocities such as racism and ethnic cleansing. Despite our intuitions at the time, most people agree that these acts are morally reprehensible. Thus, we should not rely solely on our intuitions in designing an ethical framework for self-driving cars.

Psychologist Paul Bloom has taken a similar stance against intuition, which he refers to as "empathy." He argued, "it is because of our empathetic responses that we care more about a little girl stuck in a well than about billions being affected in the future by climate change. The girl elicits empathy; statistical future harms do not" (Cook).

Recognizing that neither intuition nor conscious reasoning can stand on its own, philosopher Shelly Kagan argued that we must find a principle that satisfies both. On one hand, society wouldn't accept an ethical framework that doesn't resonate with its moral intuitions. But simultaneously, we are not striving for the most psychologically comfortable principle, but for the most morally correct principle. He discussed philosopher Frances Kamm's attempt at such principle: the Principle of Permissible Harm.

According to the principle, harm is only permissible in a case of substitution, when "the death of the one (or the deaths of the few) will be chosen—substituted—for the deaths of the greater number" (Kagan 159). The death occurs "in the very same event as the saving of a larger number of people (the greater good)" (Kagan 156). This is the case in the trolley problem where the trolley is diverted to a side track. Harm is not permissible in the case of subordination, when "we subordinate one person to another, treating the subordinate as a mere means" (Kagan 158). The death occurs due to "something that is merely a causal means to the event

that is the saving of the larger number" (Kagan 156). This is the case in the trolley problem where the large stranger is pushed off the footbridge.

Kagan argued that although Kamm's principle matches our intuitions on the trolley problem, it isn't sufficient. He stated, "the key distinction to which Kamm appeals has no obvious moral significance. When we directly consider the difference between a harm being caused by the saving of the many (or its noncausal flip side) and its being caused by a mere means to the saving of the many, it isn't at all obvious—to me, at least—why a difference like that should *matter* morally" (Kagan 158). The principle shouldn't be reverse-engineered to match our intuitions post-hoc. Instead, it should be based on a compelling rationale. So, the problem of developing an appropriate ethical framework for self-driving cars (based on trolley problems) remains unsolved.

Works Cited

Cook, Gareth. “The Moral Life of Babies.” *Scientific American*, 12 Nov. 2013, scientificamerican.com/article/the-moral-life-of-babies/.

“Jonathan Haidt.” *Ethical Systems*, NYU Stern School of Business, 2020, ethicalsystems.org/jonathan-haidt/.

Kagan, Shelly. “Solving the Trolley Problem.” *The Trolley Problem Mysteries*, by Frances Kamm, Oxford University Press, 2016, pp. 151–165.

Markie, Peter. “Rationalism vs. Empiricism.” *Stanford Encyclopedia of Philosophy*, Edited by Edward N Zalta, Metaphysics Research Lab, Stanford University, 2017.

Singer, Peter. “Ethics and Intuitions.” *The Journal of Ethics,* vol. 9, no. 3/4, 2005, pp. 331–352. *JSTOR*, jstor.org/stable/25115831.

Vickers, Salley. “Moral Tribes by Joshua Greene – Review.” *The Guardian*, Guardian News and Media, 13 Jan. 2014, theguardian.com/books/2014/jan/13/moral-tribes-joshua-greene-review.

Tragedy of the Commons

In a trolley problem scenario, should a self-driving car be able to sacrifice a passenger's life to save a greater number of people (based on utilitarian ethics)?

In a *Science* (journal) study, participants were asked this question. They answered that self-driving cars should prioritize saving the most amount of lives. However, when asked whether they would purchase this type of vehicle, participants said they wouldn't.

This makes it difficult to implement strict utilitarianism in the ethical framework of self-driving cars. By prioritizing saving the greatest number of lives, it would likely take longer for society to agree on the norms for self-driving cars. During this delay, people would still be using normal cars, so people would continue dying at a higher rate from car accidents. Thus, by trying to save more lives through utilitarianism, less lives might be saved.

This scenario is known as a "tragedy of the commons." By acting independently in the pursuit of their own interests, the collective good is spoiled. Philosophically, this study serves as a counterexample to egoism and rational egoism. For the implementation of self-driving cars, this study reveals one of the social issues that might emerge.

Few solutions have been proposed to circumvent this tragedy of the commons. In a podcast episode, Sam Harris and Paul Bloom proposed a lack of transparency as a solution. If potential customers weren't aware that their cars could sacrifice them, they might be more likely to purchase the cars.

I do not believe this is the ideal solution. People typically are interested in knowing this type of information. Some might deem it immoral for car manufacturers and regulators to withhold this information from the public. Instead, there I believe there should be a social contract. Developed by Hobbes, Locke, and Rousseau, the social contract is a model in which individuals agree to standards (e.g. moral, political, cultural norms) to exist as a society. One version of this social contract would involve people agreeing to purchase cars that followed utilitarian ethics. Alternatively, the contract could allow individuals to choose whether their cars should prioritize saving the passenger or the greatest number of lives. In this version, passengers would be legally liable if they were the passenger in a collision. This is similar to the current legal situation for non-self-driving cars, except with an added layer of nuance due to the self-driving aspect.

Works Cited

Bonnefon, Jean-François, et al. “The Social Dilemma of Autonomous Vehicles.” *Science*, vol. 352, no. 6293, American Association for the Advancement of Science (AAAS), June 2016, pp. 1573–76. DOI: 10.1126/science.aaf2654.

Cudd, Ann and Eftekhari, Seena, "Contractarianism", *Stanford Encyclopedia of Philosophy*, Edited by Edward N. Zalta, Metaphysics Research Lab, Stanford University, 2018.

Harris, Sam. “Abusing Dolores.” *Making Sense Podcast*, 12 Dec. 2016, samharris.org/podcasts/abusing-dolores.