Self Driving Cars, Values, Utilitarianism
by BRENT SILBY

Learning objectives

Objective 1	Students will be able to explain the difference between extrinsic and intrinsic value
Objective 2	Students will be able to explain why their life is of intrinsic value
Objective 3	Students will be able to explain how relinquishing their right to decide whether to sacrifice their life to an autonomous car is seen as problematic

Resource: Self Driving Cars article - Included

Content

Time	**Content**
5 minutes	**Review** Review student understanding of utilitarianism
10 minutes	**Discussion** *Perhaps run this as a think/pair/share.* Ask students what they value above anything else. They will come up with a range of answers: family, food, leisure, money. Use Socratic questioning to get them to consider that the idea that the thing they value most is their own life. e.g. why do you value those things? The students will answer in terms of other goods they bring, e.g. food brings satisfaction. Money buys products. Ask why do you value those things? Get students to see that those initial items were of instrumental value because they bring about other things of value. Try to get students to items of intrinsic value, e.g. happiness, life. Settle on one's own LIFE as something of intrinsic value which is valued above anything else.
10 minutes	**Student activity** Individually or in pairs, ask students to list situations in which they think someone else would have the right to decide whether they (the student) lives or dies. Students may come up with examples such as death penalty for terrible crime or switching off life support if there is no chance of revival. Use this as an opportunity to distinguish between legal right and moral right. In some countries the state has the legal right to decide to put someone to death. But do they have the moral right?
10 minutes	**Student reading** Students read "Self Driving Cars" article (included). Ask them to highlight points of agreement and disagreement.

Time	Content
15 minutes	**Discussion** Through class discussion, attempt to identify the argument being made in the article. Write this up on board in premise / conclusion form. Use the students' words. It should looks something like: P1. Self-driving cars will not be able to avoid all accidents P2. Self-driving cars will need to decide a course of action when they encounter life-threatening situations P3. If they are equipped with utilitarian ethics, their decisions will be designed to maximize life P4. Cars equipped with utilitarian ethics will (in life threatening situations) need to decide between protecting their driver or other people C1. Therefore Cars equipped with utilitarian ethics will sometimes need to decide to sacrifice their driver in order to protect the lives of a large number of people P5. People have the right to decide for themselves whether to sacrifice their life to protect other people. P6. Self-driving cars equipped with utilitarian ethics take away the driver's decision whether or not to sacrifice their life P7. We ought not to have our right to decide whether to sacrifice our life taken away from us. C2. Therefore we shouldn't ride in self-driving cars. Encourage students to analyse these premises. What do they agree with. Why? What do they disagree with. Why?
	Philosophical Discussion Starters Questions listed at the end of the article can serve as prompts or aids in discussion. Alternatively, they can be used in a follow up session. If used in follow up session, break students into small groups to develop answers to each of the questions. Then in full class discussion run through the questions with each group reporting back their answers.
5 minutes	**Conclusion** Summarize the lesson. Points to note: 1. extrinsic vs intrinsic value 2. life as intrinsic value 3. right to decide whether one sacrifices one's own life 4. utilitarianism weighs outcomes… 5 lives may be seen as worth more than 1

SELF DRIVING CARS

Autonomy costs autonomy

A philosophical discussion starter

By BRENT SILBY

Driving can be an exhilarating experience. The thrill of negotiating the interesting bends on a hilly country drive is a joy to many. At other times driving can be downright boring. The mind-numbingly dull drive through peak hour traffic is something most of us try to avoid. For many people, owning an autonomous vehicle – a car that can drive itself – would be something of a dream come true. No more boring traffic congestion. The car can worry about traffic while we pursue more worthy endeavors such as reading a book, watching a movie, or catching up with our social networks. It sounds promising, and many people are lining up to be among the first to own an autonomous car. However, as is often the case, benefits come at a cost. In the case of autonomous cars, the cost is a loss of driver control. For a car to be fully autonomous, the driver must relinquish some of their own autonomy – specifically, the autonomy to choose a course of action in a life-threatening situation.

Considering the number of cars on our roads, accidents are quite rare. The website gizmag.com, citing National Highway Traffic Safety Administration statistics, reveal that the rate of accidents involving property damage only is 0.38 per 100,000 miles (161,000 kilometers) driven. Despite this low number, most of us have, at some time, seen the results of an accident. And many of us have been involved in an accident. I remember driving late at night and suddenly becoming aware of an imminent collision with another car. I had a quick decision to make – do I swerve to the left or to the right? There wasn't enough time to calculate the outcome of each option, so I swerved to the right. It was a gut decision and fortunately nobody was hurt. In hindsight I think the most important thing going through my mind was avoiding the collision at any cost. The collision almost certainly would have been catastrophic, so I swerved. It was a human decision designed to ensure my own survival.

Proponents of autonomous cars might argue that such a scenario would not occur because self-driving cars do not make mistakes. They argue that autonomous cars are safer, so I would have a better chance of avoiding life-threatening accidents if I relinquish control of the wheel. This is because self-driving cars do not suffer from human failings such as fatigue, loss of attention, or poor decision making.

In principle, the possibility of a highly reliable autonomous car seems reasonable. But we are not there yet. Google cars have experienced 0.64 accidents per 100,000 miles. This is higher than the average of 0.38 accidents per 100,000 miles involving human drivers (Govers 2015). In addition to accidents, there are other cases in which Google test drivers must disengage the autonomous system. In their *Self-Driving Car Testing Report* (December 2015), Google lists the instances in which their test drivers have had to disengage the autonomous software due to some system problem or near accident. Over 424,331 miles, test drivers had to disengage the system 341 times. The most significant two reasons for disengagement were 1. perceptual discrepancy (the system became confused – perhaps seeing something that wasn't there or not seeing something that was there) and 2. software discrepancy. There were also a number of other reasons, including emergency disengagement because of the erratic behavior of another driver. Autonomous cars are clearly still vulnerable, but they are improving. In the future they will almost certainly be more reliable – especially in the ideal condition in which all cars are autonomous and share coordinate data with each other.

Let us assume that future autonomous cars will be better at making safety decisions than human drivers. Will these cars be able to avoid every conceivable accident? Probably not. I imagine they will be very good at avoiding accidents involving other autonomous cars, providing there are no external factors involved. But we live in a complex world and external factors will pose unpredictable challenges to autonomous cars. Animals sometimes run into the street. Children sometimes chase balls in front of cars. Cyclists occasionally veer

into traffic flow. How would an autonomous vehicle respond to such a scenario? It might deem the situation to be non-life threatening and continue to drive. Perhaps it would swerve to avoid the threat, as I did many years ago. On the other hand, it might decide that swerving would put too many lives at risk, in which case it would continue head on into the problem. The question as to how autonomous cars respond to life-threatening situations is where much of the controversy sits. They may be good at making safety decisions, but that does not mean they are good at making moral decisions. Essentially we are moving into a world in which we defer to our cars and ask them to make decisions as to who lives and who dies. These decisions are derived from the ethics of their programmers who do not have a direct connection to the situations the cars encounter. Because the programmers cannot be aware of the specifics of every possible scenario, they will need to develop generalized rules of action. One way to do this is to assign a value to a human life and design the car to maximize the lives saved.

A commonly cited scenario involves the car having to decide between protecting the life of the driver and protecting the lives of a group of bystanders. Suppose, for example, your autonomous vehicle is driving you down a busy road when suddenly a car full of people veers out of control into your path. Perhaps they had a tire blowout. Your car has to make a quick decision. Swerving to the left will plunge it into a sidewalk full of pedestrians, potentially killing at least five people, but keeping you alive. Swerving to the right will result in a high-speed collision into a brick wall, instantly killing you. Continuing in a straight line will mean colliding with the other vehicle, killing its passengers and you. A simple utilitarian calculation would suggest that the car ought to swerve to the right. It would kill you, but everyone else would survive. The utilitarian reasoning would be straightforward. Five lives are of more value than a single life.

Prima facie it seems reasonable for cars to be programmed to make utilitarian decisions. Indeed, many people are happy with this idea. They understand the rationale in saving as many lives as possible. But I'm not convinced. When I put my life in the hands of someone, I am essentially placing my full trust in that

person. With that trust comes the assumption that the person with whom I have placed my life will do everything possible to keep me safe. For example, when I undergo surgery, I trust the surgeon. I would never agree to surgery if I thought there was any chance that the surgeon would decide to terminate my life. I believe the same is true when I place my life in the hands of a machine. I want the machine to do everything possible to protect me.

Consider an analogous scenario. Suppose in the near future medical science progresses to the point that we have robotic surgeons. Let's assume that they are better than human surgeons because they are more precise, never tire, and have immediate access to leading edge surgical techniques. Would it be wise to put your life in the hands of a robotic surgeon? At first glance it seems like a reasonable idea. However, imagine that the robotic surgeon has been equipped with a utilitarian system which compels it to make decisions that benefit the greatest number of people. Like an autonomous car, it is compelled to save as many lives as possible. Now, imagine that during your operation, the robotic surgeon realizes that your organs are compatible with five other patients who are awaiting transplants. One of these patients needs a new heart. Another needs new lungs. The third needs a kidney, while the fourth needs a liver. The fifth patient requires a full blood transfusion. Each of these patients will die if they do not receive an immediate transplant. A simple utilitarian calculation by the robotic surgeon would lead it to conclude that your organs ought to be used to save the other five patients. You will die in the process, but the others will survive. Would you consent to undergoing a surgery with this robotic surgeon? I suspect most people would not agree to surgery if they knew there was a chance that the surgeon would decide to harvest their organs. For the same reason I think people should avoid riding in cars that are equipped with similar utilitarian protocols.

Of course, there are often instances in which people decide to commit self-sacrifice. They do so for a variety of reasons. You may one day find yourself in a situation in which you decide to sacrifice yourself for some greater good – perhaps to save the lives of a group of innocent children. But that would be *your*

decision. The problem occurs when that decision is taken away from you. As the owner of your own life, with the right to protect that life, whether or not you sacrifice your life is your decision. The same is true whether you have placed your life in the hands of another person or a computer. They should not make a decision that will lead to your death if there is another option. Implicit in the idea of placing your life in the hands of another entity is that they are to be trusted to protect you.

Autonomous cars will soon be widely available. They will undoubtedly be highly efficient and reliable. However, they will inevitably face challenging situations requiring decisions that result in loss of life. You own your own life and ought to decide for yourself whether or not you wish to sacrifice it. Allowing the autonomous car to make that decision for you means giving up your own autonomy. Regardless of how reliable these cars are, I'm not quite ready to give up my autonomy. I want to be the one who decides whether or not I sacrifice my life to protect other people. Therefore, in the meantime, I'll continue to drive for myself.

Philosophical discussion starters

1. Is a computer a rational agent?
2. Can a computer make a value judgment?
3. If a computer's values are derived from its programmer's values, do they count as the computer's values?
4. If my values are derived from my parents' values, do they count as my values?
5. Are five lives of more value than one life? Why?
6. What is the difference between instrumental and intrinsic value?
7. Does life have an intrinsic value?
8. Do I have the right to protect my own life if it means five other people lose their lives?
9. Is a harm committed if someone with whom you have placed your trust then betrays that trust for a greater good?
10. Can a computer act with good intention? Can a computer have intentions?

References

Google. (2015). *Google Self-Driving Car Testing Report on Disengagement of Autonomous Mode*, December 2015. Retrieved from https://www.google.com/selfdrivingcar/files/reports/report-annual-15.pdf

Govers, Francis. (2015). *Google reveals lessons learned (and accident count) from self-driving car program*, May 2015. Retrieved from http://www.gizmag.com/google-reveals-lessons-learned-from-self-driving-car-program/37481/