Ethical Dilemma of Self-Driving Cars: Conservative Solution

Christian Servin[1], Vladik Kreinovich[2], Shahnaz Shahbazova[3]

[1]Information Technology Systems Department,
El Paso Community College (EPCC)
919 Hunter Dr., El Paso, TX 79915-1908, USA,
cservin1@epcc.edu
[2] University of Texas at El Paso, El Paso, Texas 79968,
vladik@utep.edu
[3]Azerbaijan Technical University, Baku, Azerbaijan,
shahbazova@gmail.com

1. Self-driving cars are expected to be safer than human drivers

- Self-driving cars are supposed to provide maximum safety both:
  - for the passengers of this car and
  - for all other folks – passengers of other cars, pedestrians, and passers-by.
- In the nearest future, they are expecting to provide higher level of safety for all these categories than cars operated by human drivers.

2. Unfortunate situations, while hopefully very rare, cannot be completely avoided

- No matter how safe self-driving cars will be, unfortunate situations may still happen.
- In such situations, it may not be possible to make everyone safe.
- For example, if several pedestrians suddenly rush across the road, there may be enough time to stop the car.
- So the only choices are:
  - either hit the pedestrians
  - or swerve thus potentially hurting the car's passenger(s) and maybe even passengers of nearby cars.
- In such situations, what a car will do depends on what algorithm we program into it.
- This, in turn, depends on what objective function we use when designing this algorithm.

3. Seemingly reasonable idea: social good

- At first glance, when designing self-driving cars, we should:
  - maximize the overall social good,
  - or, equivalently, minimize the overall social harm.
- From this viewpoint:
  - if the choice is to harm (or even kill) one passenger or three pedestrians,
  - the proper solution seems to be to harm the smallest number of people,
  - i.e., in this situation, to possibly harm the passenger while trying to avoid harming the pedestrians.

4. This idea is not as reasonable as it may seem

- A detailed analysis, however, shows that such arguments may be over-simplifying and not as reasonable as they may sound at first glance.
- Let us follow one of the examples proposed by researchers.
- A medical doctor in a small town sees a reasonable healthy patient with a healthy heart, healthy liver, and two healthy kidneys.
- He/she knows that in this town, there are four patients who are:
  - at risk of dying
  - if they do not get, correspondingly, a new heart, a new liver, and a new kidney.
- Is is reasonable to kill the first patient and transplant his/her organs to the four dying folks?
- The argument is the same – shall we save the life of one patient or four patients?

5. This idea is not as reasonable as it may seem (cont-d)

- However, in this example, the answer to harm the smallest number of people does not seem so reasonable.
- To make it even less reasonable, suppose that the first patient is not fully heathy, but had a bad cut and is heavily bleeding.
- So, this patient can die if no medical help is available.
- Shall the doctor save this patient and let the other four die?
- Or shall the doctor save the lives of the four other patients by not attending to the first one?

6. So what shall we do?

- This seems like a complex problem for which we need philosophers to argue and to come up with a convincing solution.
- However, the philosophers have been discussing this "trolley problem" for many years – probably for many decades.
- And they have not yet come with a convincing solution.
- This, is to us, an indication that we should not expect such a solution in the nearest future either.
- We have to come up with such a solution ourselves.
- In this talk, we argue that such a convincing solution *is* possible.
- Namely, the solution is to be conservative and to follow the society's accepted norms and practices.

7. We must be fair to the passenger

- At present, a passenger in a car has a certain degree of safety.
- Some of this safety is provided by technical innovations such as:
  - safe and robust car design,
  - airbags, and
  - automatic warnings that inform the driver that another car is too close.
- Some of the safety is provided by the fact that the driver is in control:
  - the driver's skills – and the self-preservation instinct
  - provide safety in complex situations where technical innovations alone cannot help.

8. We must be fair to the passenger (cont-d)

- It is clearly not fair to the driver if:
  - the self-driving cars would provide a smaller degree of safety for the passenger
  - than the degree of safety obtained when this person drives the car.
- Technological progress is supposed:
  - to make all our lives better,
  - not provide advantage to some groups at the expense of others.

9. We must be fair to others

- Similarly, the self-driving cars should provide:
  - at least the same level of safety to passengers in other cars, to pedestrians, and to the passers-by,
  - as the current human-driven car.
- So:
  - if the self-driving cars focus only on the safety of their own passengers,
  - this will make it even less probable than now that the car will try to swerve to avoid hitting the pedestrian.
- In such situation, the increased safety of the passenger will come at the expense of the decreased safety for the pedestrians.

10. We must be fair to others (cont-d)

- We must be fair to pedestrians.
- We must sure that in all situations:
  - their level of safety is at least as high
  - as their current level of safety, in situations when cars are driven by human drivers.

11. The resulting idea

- This fairness is our main idea.
- Specifically, in situations when the car has the option of either harm its passenger or several pedestrians:
  - it should *not* be concerned only about the passenger – thus increasing the risk to the pedestrians, and
  - it should *not* follow the naively understood social good track idea – this increasing the risk to the passenger.
- Instead, the car should select proper probabilities of both possible actions:
  - the action that potentially hurts the passenger and
  - the action that potentially hurts the pedestrians.
- This must be done in such a way that for both groups, the level of safety be at least as high as for the current human-driven cars.

12. What should be the balance between the safety of the passenger and the safety of others

- In general, our recommendation is to make sure that:
  - the passenger is as safe as when he/she would be driving the car, and
  - others – pedestrians and bystanders – would be at least as safe as when humans drive cars.
- However, within these two restrictions, there are many possible options.
- If we are pursuing social good idea, we can keep the passenger exactly as safe as when cars are driven by people.
- In this case, we place all the efforts into minimizing the risk for others.

13. What should be the balance (cont-d)

- On the other hand:
  - if we allow customers to select which self-driving cars to buy,
  - customers will naturally want to buy a car that minimizes their risk,
  - while keeping the risk to others at the current level.
- Instead of decreasing just one of these risks – risk to the passenger and risk to others – we could try to somewhat decrease both risks.
- Which strategy should we follow? How should we balance these two risks?

14. Our idea

- We cannot solve a difficult-to-solve (and maybe even impossible-to-solve) ethical problem.
- So why not just follow what people have been doing – and what therefore is socially acceptable?
- Namely, we can find how the two risks decreased with time.
- Thus, we can find out what was, in the past, the relation between the two risks.
- This can be measured, e.g., by the percentages p_d and p_w of harmful accidents per hour of driving (or being driven) and walking.
- In general, these probabilities decrease with time; so:
  - by observing these probabilities $p_{d,i}$ and $p_{w,i}$ at different historic epochs i,
  - we can find the dependence between these two values, i.e., a function $f(p)$ for which $p_{d,i} \approx f(p_{w,i})$ for all i.

15. Our idea (cont-d)

- This function reflects a socially acceptable balance between the two risks.
- Thus, in the future:
  - when it will be possible to have self-driving cars that decrease both risks,
  - a natural idea is to use the values p_d and p_w for which $p_d = f(p_w)$.
- This will provide a socially acceptable way to balance the risks.

16. Caution

- Of course, what we propose is what medical doctors call a palliative – temporary solution that is used in lieu of a better one.
- At this moment:
  - in the absence of a better more convincing solution,
  - we propose to follow the current balance between the risks when designing self-driving cars.
- This does not mean, of course, that this conservative solution – based on the current and past social understanding – is the only way to go.
- Social norms and opinions have changed many times in the past.
- They will undoubtedly change again and again.
- What is acceptable now will no longer be acceptable.

17. Caution (cont-d)

- For example, the risk level of the original cars is not acceptable nowadays.
- So if someone wants to drive an ancient car, that car has to be retrofitted with modern safety devices.
- Maybe someone will come up with a convincing solution to the ethical dilemma.
- In all these cases, better solutions will be accepted.
- However, as of now, in the absence of such better solutions, the proposed conservative idea seems to be a reasonable way to proceed.

18. Acknowledgments

- This work was supported in part by the National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
  - HRD-1834620 and HRD-2034030 (CAHSI Includes).
- It was also supported by the AT&T Fellowship in Information Technology.
- It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.