10-1-2021

Ethical Dilemma of Self-Driving Cars: Conservative Solution

Christian Servin
El Paso Community College, cservin1@epcc.edu

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Shahnaz Shahbazova
Azerbaijan Technical University, shahbazova@gmail.com

Comments:
Technical Report: UTEP-CS-21-86

Ethical Dilemma of Self-Driving Cars: Conservative Solution

Christian Servin, Vladik Kreinovich, and Shahnaz Shahbazova

Abstract When designing software for self-driving cars, we need to make an important decision: When a self-driving car encounters an emergency situation in which either the car's passenger or an innocent pedestrian have a good change of being injured or even die, which option should it choose? This has been a subject of many years of ethical discussions – and these discussions have not yet led to a convincing solution. In this paper, we propose a "conservative" (status quo) solution that does not require making new ethical decisions – namely, we propose to limit both the risks to passengers and risks to pedestrians to their current levels, levels that exist now and are therefore acceptable to the society.

1 Formulation of a Problem

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.

Unfortunate situations, while hopefully very rare, cannot be completely avoided. No matter how safe self-driving cars will be, unfortunate situations may

Christian Servin
Information Technology Systems Department, El Paso Community College (EPCC)
919 Hunter Dr., El Paso, TX 79915-1908, USA, e-mail: cservin1@epcc.edu

Vladik Kreinovich
University of Texas at El Paso, El Paso, Texas 79968, USA
e-mail: vladik@utep.edu

Shahnaz Shabazova
Azerbaijan Technical University, Baku, Azerbaijan
e-mail: shahbazova@gmail.com

still happen, and 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 this 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, and this, in turn, depends on what objective function we use when designing this algorithm; for related discussions, see, e.g., [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30] and references therein.

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.

This idea is not as reasonable as it may seem. A detailed analysis, however, shows that such arguments may be oversimplifying and not as reasonable as they may sound at first glance. Following one of the examples proposed by researchers, suppose that a medical doctor in a small town sees a reasonable healthy patient with a healthy heart, healthy liver, and two healthy kidneys, and he/she knows that in this town, there are four patients 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? 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 the 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?

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 fact that the philosophers have been discussing this "trolley problem" for many years – probably for many decades – and have not yet come with a convincing solution 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.

What we do in this paper. In this paper, 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.

2 How to Solve the Problem: Main Idea

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, and the driver's skills – and the self-preservation instinct – provide safety in complex situations where technical innovations alone cannot help.

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.

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.

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 passenger.

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.

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 pedestrian, 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 – in such a way that for both groups, the level of safety be at least as high as for the current human-driven cars.

3 How to Solve the Problem: Details

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 by-standers – would be at least as safe as when humans drive cars. However, within these two restrictions, there are many possible options. For example:

- if we are pursuing social good idea, we can keep the passenger exactly as safe as when cars are driven by people, and place all the efforts into minimizing the risk for others;

- 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?

Our idea. Instead of trying to solve a difficult-to-solve (and maybe even impossible-to-solve) ethical problem, 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 and thus, find out what was, in the past, the relation between the two risks – as 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. 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.

Caution. Of course, what we propose is what medical doctors call a palliative – a temporary solution that is used in lieu of a better one. At this moment, in the absence of a better more convincing solution, this is what 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 moors and opinions have changed many times in the past, they will undoubtedly change again and again, and what is acceptable now will no longer be acceptable – just like the risk level of the original cars is not acceptable nowadays, and 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.

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), and 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, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

References

1. E. Barcalow, *Moral Philosophy: Theories and Issues*, Wadsworth, Belmont, CA, 2007.
2. C. W. Bauman, A. P. McGraw, D. M. Bartels, and C. Warren, "Revisiting external validity: concerns about trolley problems and other sacrificial dilemmas in moral psychology", *Social and Personality Psychology Compass*, 2014, Vol. 8, No. 9, pp. 536–554.
3. R. M. Bernhard, J. Chaponis, R. Siburian, P. Gallagher, K. Ransohoff, D. Wikler, H. Roy, and J. D. Greene, "Variation in the oxytocin receptor gene (OXTR) is associated with differences in moral judgment", *Social Cognitive and Affective Neuroscience*, 2016, Vol. 11, No. 12, pp. 1872–1881.
4. A. Bleske-Rechek, L. A. Nelson, J. P. Baker, M. W. Remiker, and S. J. Brandt, "Evolution and the trolley problem: People save five over one unless the one is young, genetically related, or a romantic partner", *Journal of Social, Evolutionary, and Cultural Psychology*, 2010, Vol. 4, No. 3, pp. 115–127.
5. J.-F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of autonomous vehicles", *Science*, 2016, Vol. 352, No. 6293, pp. 1573–1576.
6. E. Ciaramelli, M. Muccioli, E. Làdavas and G. di Pellegrino, "Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex", *Social Cognitive and Affective Neuroscience*, 2007, Vol. 2, No. 2, pp. 84–92.
7. M. J. Crockett, L. Clark, M. D. Hauser, and T. W. Robbins, "Serotonin selectively influences moral judgment and behavior through effects on harm aversion", *Proceedings of the National Academy of Sciences*, 2010, Vol. 107, No. 40, pp. 17433–17438.
8. E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The Moral Machine experiment", *Nature*, 2018, Vol. 563, No. 7729, pp. 59–64.
9. H. B. Francis, C. Howard, I. S. Howard, M. Gummerum, G. Ganis, G. Anderson, and S. Terbeck, "Virtual morality: transitioning from moral judgment to moral action?", *PLOS One*, 2016, Vol. 11, No. 10, pp. 1–22.
10. J. Gogoll, J. F. Müller, and F. Julian, "Autonomous cars: in favor of a mandatory ethics setting", *Science and Engineering Ethics*, 2017, Vol. 23, No. 3, pp. 681-–700.
11. J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen, "An fMRI investigation of emotional engagement in moral judgment", *Science*, 2001, Vol. 293, No. 5537, pp. 2105–2108.
12. J. Himmelreich, "Never mind the trolley: the ethics of autonomous vehicles in mundane situations", *Ethical Theory and Moral Practice*, 2018, Vol. 21, No. 3, pp. 669–684.
13. N. JafariNaimi, "Our bodies in the trolley's path, or why self-driving cars must *not* be programmed to kill", *Science, Technology, and Human Values*, 2018, Vol. 43, No. 2, pp. 302–323.
14. J. Jarvis Thomson, "Killing, letting die, and the trolley problem", *The Monist*, 1976, Vol. 59, pp. 204–217.
15. J. Jarvis Thomson, "The trolley problem", *Yale Law Journal*, 1985, Vol. 94, No. 6, pp. 1395–1415.
16. G. Kahane, "Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment", *Social Neuroscience*, 2015, Vol. 10, No. 5, pp. 551–560.
17. G. Kahane, J. A. C. Everett, B. D. Earp, L. Caviola, N. S. Faber, M. J. Crockett, and J. Savulescu, "Beyond sacrificial harm: A two-dimensional model of utilitarian psychology", *Psychological Review*, 2018, Vol. 125, No. 2, pp. 131–164.

18. M. Lee, S. Sul, and H. Kim, "Social observation increases deontological judgments in moral dilemmas" *Evolution and Human Behavior*, 2019, Vol. 39, No. 6, pp. 611–621.
19. H. S. M. Lim and A. Taeihagh, "Algorithmic decision-making in AVs: understanding ethical and technical concerns for smart cities", *Sustainability*, 2019, Vol. 11, No. 20, Paper 5791.
20. F. Myrna Kamm, "Harming some to save others", *Philosophical Studies*, 1989, Vol. 57, No. 3, pp. 227–260.
21. C. D. Navarrete, M. M. McDonald, M. L. Mott, and B. Asher, "Virtual morality: Emotion and action in a simulated three-dimensional 'trolley problem' ", *Emotion*, 2021, Vol. 12, No. 2, pp. 364–370.
22. I. Patil, C. Cogoni, N. Zangrando, L. Chittaro, and G. Silani, "Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas", *Social Neuroscience*, 2014, Vol. 9, No. 1, pp. 94—107.
23. S. Rom and C. Paul, "The strategic moral self: self-presentation shapes moral dilemma judgments", *Journal of Experimental Social Psychology*, 2017, No. 74, pp. 24–37.
24. F. C. Sharp, "A Study of the influence of custom on the moral judgment", *Bulletin of the University of Wisconsin*, 1908, No. 236, p. 138.
25. F. C. Sharp, *Ethics*, The Century Co., New York, 1928.
26. A. Skulmowski, A. Bunge, K. Kaspar, and G. Pipa, "Forced-choice decision-making in modified trolley dilemma situations: a virtual reality and eye tracking study", *Frontiers in Behavioral Neuroscience*, 2014, Vol. 8, No. 426, pp. 1–16.
27. L. R. Sütfeld, R. Gast, P. König, and G. Pipa, "Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios: Applicability of Value-of-Life-Based Models and Influences of Time Pressure", 2017, *Frontiers in Behavioral Neuroscience*, Vol.11, No. 122, pp. 1–13.
28. P. Unger, *Living High and Letting Die*, Oxford University Press, Oxford, UK, 1996.
29. P. Valdesolo and D. DeSteno, "Manipulations of Emotional Context Shape Moral Judgment", *Psychological Science*, 2006, Vol. 17, No.6, pp. 476–477.
30. F. F. Yousseff, K. Dookeeram, V. Basdeo, E. Francis, M. Doman, D. Mamed, S. Maloo, J. Degannes, and L. Dobo, "Stress alters personal moral decision making", *Psychoneuroendocrinology*, 2012, Vol. 37, No. 4, pp. 491–498.