4-1-2024

How to Fairly Allocate Safety Benefits of Self-Driving Cars

Fernando Munoz
The University of Texas at El Paso, fmunoz9@miners.utep.edu

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

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

Comments:
Technical Report: UTEP-CS-24-19

How to Fairly Allocate Safety Benefits of Self-Driving Cars

Fernando Muñoz, Christian Servin, and Vladik Kreinovich

Abstract In this paper, we describe how to fairly allocated safety benefits of self-driving cars between drivers and pedestrians – so as to minimize the overall harm.

1 Formulation of the Practical Problem

Our roads are still not perfectly safe. In spite of all the efforts of transportation engineers, there are still about 6 million car accidents every year in the US only.

How can we make roads safer: the promise of self-driving cars. Most road accidents are caused by human errors – which, in their turn, are caused by fatigue, distractions, etc. Because of this, a natural way to decrease this number of accidents and to make roads safer is to replace some – or even all – human decision making with decision made by automated systems. The current self-driving cars have not yet reached their safety potential, but they are getting better, and we all hope that in the nearest future they will be safer to drive than the usual human-driven vehicles.

Resulting problem. Self-driving cars will (hopefully) bring more safety. When an automatic system is in charge, a system that does not know fatigue, that is never under the influence, that does not violate traffic rules, there will be hopefully much fewer accidents, and most of the remaining accidents will be less severe than they

Fernando Muñoz
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: fmunoz9@miners.utep.edu

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

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: vladik@utep.edu

are now – since the reaction time of an automatic system is much smaller than the reaction time of a human driver.

It would be great to eliminate all accidents, but we do not think that this will happen in the nearest future: while some cars will be self-driving, many other cars will remain driven by humans who are more error-prone, not to mention that the roads are also shared by pedestrians who do not always obey the traffic rules.

In pedestrian-car situations when an accident is inevitable and someone will get hurt, the automatic system often has a choice. For example, it can swerve into the wall thus injuring the car and the driver but leaving the pedestrian intact, or is can try to make the driver safer but risking hitting and injuring the pedestrian. In human-driven cars, this decision is left to the driver (and the car manufacturers try to minimize the impact both on the driver and on the pedestrian by designing safer cars and cars with bumpers that bring the least harm to pedestrians). However, in a self-driving car, this decision needs to be made by the automatic system; see, e.g., [1] and references therein.

How to allocate safety benefits: a practical problem. What this system will do depends on how we program it.

What we proposed earlier. In our previous paper [1], we proposed to follow social norms when designing the system. So what decision should we program?

What we would prefer to do. Instead of relying on social norms – which do not always perfectly describe what is the best for the society – it is desirable to select such programs for which the overall harm to the society – i.e., the overall harm to pedestrians and to drivers – is minimized.

What we do in this paper. In this paper, we propose a solution to this problem.

The structure of the paper. We start, in Section 2, with a simple seemingly adequate model of the situation. In Section 3, we explain why this – as it turned out, oversimplified – model is not fully adequate for this problem, and what else needs to be taken into account to come up with an adequate model. In Section 4, we describe our final model. In Section 5, we show that this model leads to a simple straightforward algorithm for the optimal allocation of safety benefits.

2 First-approximation model

What causes pedestrian-car accidents? Our goal is to minimize the overall expected harm caused by pedestrian-car accidents. To gauge the related harm, let us first recall what causes these accidents.

- Some incidents are caused by drunk drivers – this problem will be completely eliminated if we switch to self-driving cars.
- Some incidents are caused by an accident: a person absent-mindedly walks into a street, is accidentally pushed into the traffic area, etc. Such accidents are rela-

tively rare. Let us denote by n_a the proportion of such accidents per single street crossing.
- The vast majority of car incidents involving pedestrians occur when a pedestrian tries to cross a street at the wrong place – or at red light. The ability of cars to self-drive will, unfortunately, not cause pedestrians to be more disciplined – so this needs to be taken into account.

How to gauge the average harm caused by pedestrian-car accidents: current situation. Let us estimate the average harm per one street crossing. In view of the above, to estimate this average harm, we need to know:

- what is the proportion of street crossings that are not following the rules; let us denote this proportion by n;
- what is the proportion of not-following-the-rules crossings that result in an accident; let us denote this proportion by a; and
- what is the average harm done to the pedestrian and to the driver; let us denoted these harms by, correspondingly, h_p and h_d.

In these terms, the average harm per street crossing is equal to the product $n \cdot a \cdot (h_p + h_d)$.

What will happen when we use self-driving cars. When we introduce effective self-driving cars, the following will happen:

- some proportion p_+ of accidents will be avoided completely;
- some proportion p_- of accidents we will still not be able to avoid at all; and
- the remaining proportion $p_0 = 1 - p_+ - p_-$ is when we can decrease the harm, but not fully eliminate it.

It is in this last category, in which we cannot fully eliminate the harm, that we need to decide how to fairly distribute the benefits of increase safety.

Let $b_p \in [0,1]$ indicate the proportion of benefits that goes to the pedestrian, then $b_d = 1 - b_p$ will be the proportion of benefits that goes to the driver. In these terms:

- the expected harm to the pedestrian will be equal to

$$(n_a + n) \cdot a \cdot h_p \cdot (p_- + p_0 \cdot (1 - b_p)), \text{ and}$$

- the expected harm to the driver will be equal to

$$(n_a + n) \cdot a \cdot h_d \cdot (p_- + p_0 \cdot (1 - b_d)).$$

Thus, the overall expected harm is equal to the sum of these harms:

$$(n_a + n) \cdot a \cdot h_p \cdot (p_- + p_0 \cdot (1 - b_p)) + (n_a + n) \cdot a \cdot h_d \cdot (p_- + p_0 \cdot (1 - b_d)). \quad (1)$$

3 Limitations of this model

What does this model recommend? At first glance, our resulting formula (1) provides a reasonable description of the average harm. So, all we need to do now is to find the value b_p that minimizes the average harm.

To find this optimal value, let us plug in the expression $b_d = 1 - b_p$ into this formula. In this case, $1 - b_d = b_p$ and thus, we get the following expression:

$$(n_a + n) \cdot a \cdot h_p \cdot (p_- + p_0 \cdot (1 - b_p)) + (n_a + n) \cdot a \cdot h_d \cdot (p_- + p_0 \cdot b_p). \qquad (2)$$

We get a linear expression in terms of the unknown b_p. Separating terms depending and not depending on b_p in this expression, we can transform the expression (2) into the following form:

$$\text{const} + b_p \cdot (n_a + n) \cdot a \cdot p_0 \cdot (h_d - h_p), \qquad (3)$$

where by const, we mean terms that do not depend on b_p at all.

In a pedestrian-car accident, the harm to the pedestrian is usually much larger than the harm to the driver: $h_p \gg h_d$. Thus, to minimize the expression (3) that describes the overall harm, we need to maximize the benefits b_p allocated to the pedestrian, i.e., to select $b_p = 1$.

What is wrong with this recommendation? What will happen if we follow this recommendation? Let us consider the extreme case when $p_- = 0$, i.e., when the effect of *all* accidents will be decreases. In this case, if we select $b_p = 1$, this means that even when a pedestrian crosses the street in the wrong place, he/she will not harmed: the only person who may be harmed will be the driver.

In our cities, we already have a lot of people jaywalking – even those sometimes this causes them harm. If we decrease the risk to pedestrians, what will prevent even more people to start jaywalking and thus, increasing the number of accidents even more?

This is clearly not what we expected.

So what was wrong with the model? Since the recommendations provided by our model are not fully adequate, this means that something in this model was not adequate.

The above reasoning explains what was wrong – we implicitly assumed that the proportion n of people who do not follow the rules when crossing the street is a constant. In reality, as the expected harm decreases, this proportion will increase. To make our model more adequate, we need to take this increase into account.

4 Final model

How can we determine the proportion n of pedestrians who do not follow the rules? To answer this question, let us recall why pedestrians sometimes do not fol-

low the rules: because if they do not wait for the green light and/or do not walk to the official crossing, they gain some time. Let us denote the average gain per each crossing by g.

Each pedestrian has a choice:

- either to follow the rules and thus, lose the potential gain amount g,
- or not to follow the rules and thus, face the loss $n \cdot a \cdot h_p \cdot (p_- + p_0 \cdot (1 - b_p))$.

If the second loss is smaller, more pedestrians will select not following the rules and thus, the proportion n will increase – until these numbers even out. Similarly, if the first loss is smaller, more pedestrians will decide to follow the rules, and thus, the proportion n will decrease – until these numbers even out. Thus, eventually, these losses will become equal:

$$g = n \cdot a \cdot h_p \cdot (p_- + p_0 \cdot (1 - b_p)).$$

Based on this equality, we can find the desired proportion n:

$$n = \frac{g}{a \cdot h_p \cdot (p_- + p_0 \cdot (1 - b_p))}. \tag{4}$$

Our final model. Substituting the expression (4) into the formula (2), we get our final model, that describes the expected average harm corresponding to the given value b_p:

$$\left(n_a + \frac{g}{a \cdot h_p \cdot (p_- + p_0 \cdot (1 - b_p))}\right) \cdot H, \text{ where}$$

$$H \stackrel{\text{def}}{=} a \cdot h_p \cdot (p_- + p_0 \cdot (1 - b_p)) + a \cdot h_d \cdot (p_- + p_0 \cdot b_p). \tag{5}$$

To find the best allocation of safety benefits between pedestrians and drivers, we need to select the value b_p from the interval $[0, 1]$ that minimizes the expression (5).

Why we believe that our new model is more adequate. We dismissed our original simplified model because it led to an inadequate solution $b_p = 1$. Let us show that our new, more adequate model does not have this problem, at least when the value p_- – that describes that remaining imperfection of the self-driving cars – is sufficiently small.

Indeed, in the case of $p_- = 0$, when b_p tends to 1, the first factor in the formula (5) tends to infinity, while the second factor remains non-zero. Thus, the product (5) also tends to infinity when b_p tends to 1. So, in contrast to the above simplified model, the minimum of the expression (5) cannot be attained when $b_p = 1$.

5 How to find the optimal allocation b_p

Analysis of the problem. How can we actually find the optimal allocation b_p for which the expression (5) – that describes the expected average harm - attains its smallest possible value? Good news is that we can actually find an explicit expression for this optimal value. Let us explain how to do it.

The dependence of the expression (5) on the unknown b_p has the following form:

$$\left(a_0 + \frac{a_1}{a_2 + a_3 \cdot b_p}\right) \cdot (a_4 + a_5 \cdot b_p), \tag{6}$$

for the following coefficients a_i:

$$a_0 = n_a, \;\; a_1 = g, \;\; a_2 = a \cdot h_p \cdot (p_- + p_0), \tag{7}$$

$$a_3 = -a \cdot h_p \cdot p_0, \;\; a_4 = a \cdot h_p \cdot (p_- + p_0) + a \cdot h_d \cdot p_i, \;\; a_5 = -a \cdot h_p \cdot p_0 + h_d \cdot \tag{8}$$

According to calculus, the minimum of a differentiable function inside an interval $b_p \in (0,1)$ is attained at a point where the derivative of this function is equal to 0. Differentiating the expression (6) with respect to b_p and equating the derivative to 0, we get the following equation:

$$-\frac{a_1 \cdot a_3}{(a_2 + a_3 \cdot b_p)^2} \cdot (a_4 + a_5 \cdot b_p) + \left(a_0 + \frac{a_1}{a_2 + a_3 \cdot b_p}\right) \cdot a_5 = 0. \tag{9}$$

Multiplying both sides of this equation by $(a_2 + a_3 \cdot b_p)^2$, we get the following quadratic equation:

$$-a_1 \cdot a_3 \cdot (a_4 + a_5 \cdot b_p) + a_0 \cdot a_5 \cdot (a_2 + a_3 \cdot b_p)^2 + a_1 \cdot a_5 \cdot (a_2 + a_3 \cdot b_p) = 0,$$

i.e., the equation

$$A \cdot b_p^2 + B \cdot b_p + C = 0, \tag{10}$$

where we denoted

$$A = a_0 \cdot a_5 \cdot a_3^2, \tag{11}$$

$$B = -a_1 \cdot a_3 \cdot a_5 + 2a_0 \cdot a_5 \cdot a_2 \cdot a_3 + a_1 \cdot a_5 \cdot a_3 = 2a_0 \cdot a_5 \cdot a_2 \cdot a_3, \tag{12}$$

$$C = -a_1 \cdot a_3 \cdot a_4 + a_0 \cdot a_5 \cdot a_2^2 + a_1 \cdot a_5 \cdot a_2. \tag{13}$$

Now, we can use the general formula for solving a quadratic equation to find the optimal value of b_p:

$$b_p = \frac{-B \pm \sqrt{B^2 - 4A \cdot C}}{2A}. \tag{14}$$

Out of two possible solutions, we should select the one that is inside the interval $(0,1)$ – or, if both solutions are inside this interval, the one for which the value of the expression (5) is the smallest.

So, we arrive at the following algorithm.

Algorithm for computing the parameter b_p that describes optimal allocation of safety benefits of self-driving cars. To compute b_p, we need to know the following values:

- the proportion a of not-following the rules crossings that currently lead to an accident;
- the proportion n_a of crossings that result in accidental violation of the rules;
- the current average per-accident harm h_p to a pedestrian;
- the current average per-accident harm h_d to a driver;
- the proportion p_+ of the accidents that will be completely avoided by the self-driving cars;
- the proportion p_- of the accidents about which the self-driving car system cannot do anything.

Based on these values, we can compute the proportion $p_0 = 1 - p_+ - p_-$ of the accidents for which the self-driving car system can decrease – but not fully eliminate – the harm. What we want to compute is the proportion b_p of the resulting safety benefits that will be allocated to pedestrians. To compute the optimal proportion b_p, we do the following:

- first, we use the formulas (7)–(8) to compute the auxiliary coefficients a_i;
- then, we use the formulas (11)–(13) to compute the auxiliary coefficients A, B, and C;
- finally, we use the formula (14) to compute the optimal value of b_p.

Out of two possible solutions, we should select the one that is inside the interval $(0,1)$ – or, if both solutions are inside this interval, the one for which the value of the expression (5) is the smallest.

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), HRD-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES), and by the AT&T Fellowship in Information Technology.

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

References

1. C. Servin, V. Kreinovich, and S. Shahbazova, "Ethical dilemma of self-driving cars: conservative solution", In: S. N. Shahbazova, A. M. Abbasov, V. Kreinovich, J. Kacprzyk, and I. Batyr-

shin (eds.), *Recent Developments and the New Directions of Research, Foundations, and Applications*, Springer, Cham, Switzerland, 2023, Vol. 2, pp. 93–98.