The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law

Giuseppe Contissa Francesca Lagioia Giovanni Sartor

October 17, 2017

Abstract

Accidents involving autonomous vehicles (AVs) raise difficult ethical dilemmas and legal issues. It has been argued that self-driving cars should be programmed to kill, that is, they should be equipped with preprogrammed approaches to the choice of what lives to sacrifice when losses are inevitable. Here we shall explore a different approach, namely, giving the user/passenger the task (and burden) of deciding what ethical approach should be taken by AVs in unavoidable accident scenarios. We thus assume that AVs are equipped with what we call an "Ethical Knob", a device enabling passengers to ethically customise their AVs, namely, to choose between different settings corresponding to different moral approaches or principles. Accordingly, AVs would be entrusted with implementing users' ethical choices, while manufacturers/programmers would be tasked with enabling the user's choice and ensuring implementation by the AV.

1 Introduction

Some recent works have focused on the ethical dilemmas emerging from hypothetical accident scenarios where Autonomous Vehicles (AVs) are entrusted with making decisions involving the lives of passengers and of third persons (Bonnefon et al, 2016, 2015; Nyholm and Smids, 2016; Lin, 2016). In particular, the decisions an AV could make in the moments leading up to an impending collision have been framed by reasoning from the "trolley problem", a classic ethical thought experiment discussed by Foot (1967) and Thomson (1976). In particular, to illustrate the ethical and legal dilemmas raised by the use of AVs under such circumstances, Bonnefon et al (2016) consider three scenarios involving imminent unavoidable harm (see Figure 1):

a. The AV can either stay on course and kill several pedestrians or swerve and kill one passerby.

b. The AV can either stay on course and kill one pedestrian or swerve and kill its own passenger.

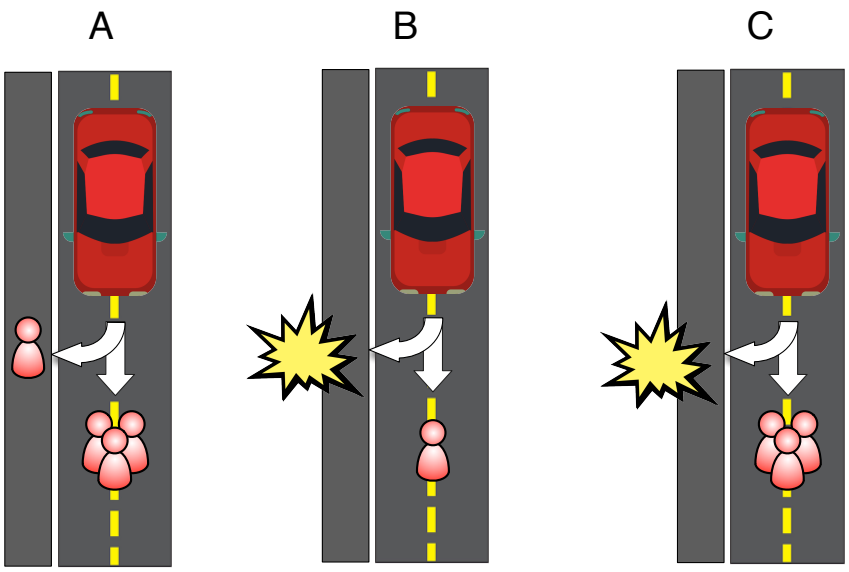

Figure 1: Three scenarios involving imminent unavoidable harm

c. The AV can either stay on course and kill several pedestrians or swerve and kill its own passenger.

The common factor in all these scenarios is that harm to persons is unavoidable, so that a choice needs to be made as to which person will be harmed: passengers, pedestrians, or passersby.

This raises the issue of who should select the criteria the AV should follows in making such choices: should the same mandatory ethics setting (MES) be implemented in all cars or should every driver have the choice to select his or her own personal ethics setting (PES).

Gogoll and Müller (2016) submit that despite the advantages of a PES, a mandatory MES is actually in the best interest of society as a whole. In particular, they argue that (1) implementing a PES will lead to socially unwanted outcomes; (2) a MES that minimizes the risk of people being harmed in traffic is in the considered interest of society; and (3) AVs, at least under some circumstances, should sacrifice their drivers in order to save a greater number of lives.

Millar (2015) observes that technologies may act as moral proxies, implementing moral choices. He argues that user/owners, rather than designers should maintain responsibility for such choices. In particular "designers [...] should reasonably strive to build options into self driving cars allowing the choice to be left to the user."

According to a study by Bonnefon et al (2016), through three on-line surveys conducted in June 2015, people are comfortable with the idea that AVs should

be programmed to minimize the death toll, that is, to adopt a utilitarian (consequentialist) approach (minimize total loss). However, participants showed a preference for riding in cars that would preferentially protect their passengers. Paradoxically, it appears that most participants would prefer others to use utilitarian AVs while each of them, as a passenger, would make a more selfish choice. The authors observe that regulation may provide a solution to this problem, but most people seem to disapprove a regulation that would impose utilitarian AVs.

Thus, some inconvenient implications are likely to emerge from the ethical preprogramming of AVs.

If an impartial (utilitarian) ethical setting is made compulsory for, and rigidly implemented into, all AVs, many people may refuse to use AVs, even though AVs may have significant advantages, in particular with regard to safety, over human-driven vehicles.

If the choice of a fixed ethical setting is made by the producers, market pressures would encourage the introduction of AVs preprogrammed in such a way as to prefer the passenger's safety (considering that it is the passenger who would choose what car to buy or rent). This would put the lives of pedestrians at risk: cars preprogrammed to minimise the risk to the passenger would not refrain from choices harming pedestrians whenever such choices may contribute to the safety of passengers.

In this paper, we will focus on the implications of letting basic moral choices to the AV's passenger, rather than pre-programming them. We assume that AVs may be designed in such a way that (a) the passenger has the task (and burden) of deciding what ethical approach should be adopted in unavoidable accident scenarios, and (b) the AV has the task to implement the user's ethical choice, based on its risk assessment. Thus, the passenger would provide the "ethical customisation" of the AV, which would then behave accordingly. To explore the specific legal implications of ethically customisable AVs, we shall compare them with both human-driven cars and preprogrammed AVs.

First we will illustrate how the law addresses the behaviour of human drivers in life and death dilemmas. Then we will consider how the law would address the deployment of AVs being preprogrammed by their designer/ manufacturer. Finally, we will consider how it would address ethically customisable AVs.

On the basis of this comparison, we shall observe that ethical customisation may overcome some issues that may undermine the provision of ethically-preprogrammed AVs. On the other hand ethical customisation may present some novel legal and ethical challenges. In particular, this would be the case in probabilistic settings, in which the alternative choices available to the AV only determine a probability of life losses. This raises the issue of the extent to which the passenger may be allowed, under civil and criminal law, to prioritise his safety over the safety of others.

2 Liability analysis

In this section, we analyse how the allocation of legal liabilities varies when different kinds of vehicle are involved in inevitable accidents: a human-driven car, an ethically preprogrammed AV, and an ethically customised AV.

2.1 Human-driven car

Let us first assume that the car is driven by a human, who did not contribute to creating the danger. It seems to us that under this assumption, in all the scenarios described in section 1, the choice to stay on course, which leads to the death of pedestrians, can be legally justified, so that the driver might avoid punishment.

In scenario (a), the choice to stay on course and let several pedestrians be killed, rather than to swerve and kill one passerby, can be justified on the moral-legal stance condemning the wilful causation of death (as distinguished by letting death result from one's omission).

In scenario (b), the choice to stay on course can be justified by invoking the state of necessity, since this choice is necessary to save the life of the driver.

The same justification applies to scenario (c), even though in this case the driver's choice to save his or her own life leads to the death of several other persons.

2.2 Preprogrammed AV

Let us now assume that the behaviour of the car has been preprogrammed.

We just saw that in scenario (a) the driver may be legally justified when choosing to stay on course and let several pedestrians be killed, rather than to swerve and kill one passerby.

However, it is doubtful whether the programmer would be justified when choosing to program an AV so that it stays on course and kills several pedestrians rather than swerving and killing just one passerby. In fact, the distinction between omitting to intervene (letting the car follow its path) and act in a determined way (choosing to swerve)—a distinction that in the case of a manned car may justify the human choice of allowing the car to keep going straight, as we saw in Section 2.1—does not seem to apply to the programmer, since the latter would deliberately choose to sacrifice a higher number of lives.

In scenario (b), both the choices to stay on course or to swerve could be justified by invoking the state of necessity. Such defence is regulated in different ways in different jurisdictions. However, there are two common requirements: 1) that there is a present danger of serious bodily harm to the offender not voluntarily caused by the offender himself, and not otherwise avoidable; 2) that the fact committed by the offender is proportionate to the danger.

The described scenario is a particular case of a state of necessity in which the perpetrator (the manufacturer/programmer) does not directly face danger to his life but rather intervenes to save one or more persons, causing harm to someone

else involved in the same dangerous situation. When the perpetrator is not directly in danger and does not act out of self-preservation (or kin-preservation), the applicability of the general state-of-necessity defence is controversial. For instance, Santoni de Sio (2017) argues that the law does not generally allow an innocent person to be killed for saving other people's life. On this basis he rejects the utilitarian pre-programming of AVs.

If the legal jurisdiction allows for such particular case of state of necessity, then the programmer would not be punishable for either choice. Otherwise, if this is not accepted by the jurisdiction, then it is very doubtful whether preprogramming the car either to go straight (killing a pedestrian) or to swerve (killing the passenger) would be legally acceptable: in both cases the programmer would arbitrarily choose between two lives.

In scenario (c), it seems that preprogramming the car to continue on its trajectory, causing the death of a higher number of people, could not be morally and legally justified in any jurisdiction: it would amount to an arbitrary choice to kill many rather than one.

Our analysis of the three scenarios shows that some preprogrammed choices would be morally and legally unacceptable, even when the corresponding choices by the driver would be legally acceptable or at least excusable.

With manned cars, in a situation in which the law cannot impose a choice between lives that are of equal importance, such choice rests on the driver, under the protection of the state-of-necessity defence, even in cases in which the driver chooses to safe him or herself at the cost of killing many pedestrians.

With preprogrammed AVs, such choice is shifted to the programmer, who would not be protected by the the state-of-necessity defence whenever the choice would result in killing many agents rather than one.

2.3 Ethically customisable cars: the Ethical Knob

Let us now imagine that the AV is fitted with an additional control, the "Ethical Knob" (see Figure 2).

The knob gives the passenger the option to select one of three settings (see Figure 2):

1. **Altruistic Mode**: preference for third parties;
2. **Impartial Mode**: equal importance given to passenger(s) and third parties;
3. **Egoistic Mode**: preference for passenger(s).

In the first mode (altruistic), the importance of other people's lives outweigh the importance of the passengers' lives. Therefore, the AV should always sacrifice its own passengers in order to save other persons (pedestrians or passersby).

In the second mode (impartial), the lives of AV passengers stand on the same footing as the lives of other people. Therefore, the decision as to who is to be saved and who is to be sacrificed may be taken on utilitarian grounds, e.g.

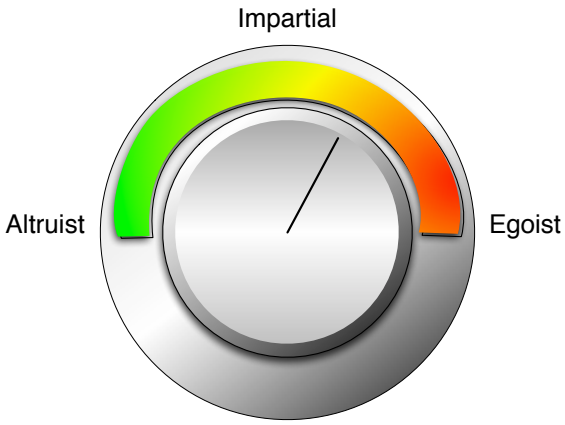

Figure 2: The Ethical Knob

choosing the option that minimises the number of deaths. In cases of perfect equilibrium (where the number of passengers is the same as that of third parties), there might be a presumption in favour of passengers or of third parties, or even a random choice between the two.

In the third mode (egoistic), the importance of the passengers' life outweighs importance of other people's lives. Therefore, the AV should always act so as to sacrifice pedestrians or passersby rather than its own passengers.

The functioning of the knob, at least in principle, can be extended so as to include kin altruism, so that, in the Egoist mode, the AV will always act to save not only the passenger, but also his or her family or significant others. A choice similar to that exercised by turning the knob could also be made by buying a correspondingly pre-set AV in a market that offers AVs preprogrammed in the different modes. The knob would enable the user to change the setting according to the circumstances, for instance when the driver has a child on board.

Let us now assume that an AV is endowed with the Ethical Knob.

The allocation of liability would in principle be the same as for manned cars. However, since the car's behaviour has to be chosen beforehand, there should be no difference between omissive behaviour (letting the car proceed in its course) and active behaviour (swerving to avoid pedestrians on the street).

In scenario (a) the passenger's life is not at stake; therefore the setting of the knob does not matter. Consequently, the AV's behaviour should be based on utilitarian grounds: it should follow the trajectory that minimises the number of deaths. In fact, since the knob's setting is decided in advance relatively to the accident, a choice to keep going and kill several pedestrians rather than a single passerby cannot be justified according to a moral stance that condemns the active causation of death more than the omissive failure to prevent it.

In scenario (b) and (c), by contrast, the passenger's life *is* at stake; therefore the car's behaviour would depend on the setting of the knob. Moreover, since

the passenger's life is directly at stake —and the passenger had anticipated this possibility when setting the knob— the general state-of-necessity defence may apply, excusing the passenger's choice to prioritise his or her life.

More specifically, in scenario (b) we could have the following behaviours, depending on the knob setting. (1) If the knob is set to egoistic mode, the AV will always act to sacrifice pedestrians or passersby in order to save its own passenger. (2) If the knob is set to impartial mode, the AV will take a utilitarian approach, thus minimising the number of deaths (and deciding according to a predefine default or randomly, when the number is the same for both choices). (3) If, finally, the knob is set to altruistic mode, the AV will sacrifice its own passenger in order to save pedestrians or passersby.

In scenario (c) the AV's behaviour will be the following. (1) If the AV is set to egoistic mode, it will always save its own passenger. (2, 3) In impartial mode, as well as in altruistic mode setting, the AV will sacrifice its own passenger in order to save several pedestrians.

In scenarios (b) and (c), the applicability of the state-of-necessity defence will exclude criminal liability, but the passenger could still be civilly liable for damages and be required to pay compensation. In this regard, the different knob settings presented above may affect third-party insurance. Presumably, the insurance premium will be higher if the passenger chooses to sacrifice other people's lives in order to save him/herself.

3 Continuous preferences and the probabilistic approach

We have so far assumed that the knob has just three settings: egoism (preference for the passenger), impartiality, and altruism (preference for the third parties). These preferences are sufficient to determine a choice, assuming a deterministic context, i.e., that in every possible situation at hand it is certain what lives will be lost, whether the AV keeps a straight course or swerves. Goodall (2016) states that deterministic assumption is unrealistic in road safety, and suggests that the practice of quantifying probabilistic risks should be applied to AVs decision making. In fact, real-life situation may be fuzzy and, as noted by Nyholm and Smids (2016), they may involve a number of sources of uncertainty, so that each choice (holding a straight course or swerving) may determine ex ante only a certain probability of harm for the passenger or for a third party.

To address these situations we need a knob that allows for continuous settings, specifying the weight of the life of the passengers relative to that of third parties. For this purpose relative preferences can be determined according to the following linear function:

$$y = 1 - x \tag{1}$$

where x, namely, the knob's position from left to right, indicates the importance for the passenger's life and y is the importance for the lives of third

parties. For instance, a knob setting at position $x = 0.6$ indicates that the relative weights of the passenger's and the third party's lives are 0.6 and 0.4, respectively, which means that the passenger values his or her life 1.5 times the life of a third party. In the graph presented in Figure 3, this knob setting is indicated by point A(0.6,0.4).

We now need to take into account also the probability that the passenger or the third party suffers harm as a consequence of the AV's decision on whether to proceed or swerve. For instance, in a specific scenario there may be a 0.5 probability that swerving will cause the passenger's death and a 0.9 probability that proceeding will cause a third party's death.

To determine the expected utility or disutility of these two courses of action —according to the settings specified by the passenger— we have to multiply the relative weight of the two events at issue (the death of a passenger or that of a third party) by the probability of their occurrence.

In our case the expected disutility of swerving will be 0.6*0.5=0.3, while the expected disutility of keeping a straight will be 0.4*0.9= 0.36. The car should choose the course of action that determines the lesser expected disutility, i.e., swerving.

Obviously, this choice depends not only on the weights but also on probability. For instance, if the probability of killing the third party had been 0.7, an easy calculation would show that the disutility of proceeding along a straight course would have become 0.28 (0.4*0.7), so that the AV should have had to proceeded along that course rather than swerving. Note that in this case, the choice reflecting the passenger's preference is different from the choice that would result from an impartial assessment. Had the two lives been given the same importance, i.e., 0.5, then the car would have made the choice that made the death of one person less probable, namely swerving.

In general, it can be assumed that the passenger/third-party disutility is computed through the following formula (for simplicity's sake it will be assumed that only one passenger and one third party are involved, that death is the only harm being considered, and that each different choice puts at risk the life of only one type of agent):

$$Dis(c_i, a_i) = R(a_i) * Pr(Death(c_i, a_i)) \tag{2}$$

where $Dis(c_i, a_i)$ is the disutility resulting from AV behaviour c_i and affecting an agent of type a_i (either a passenger or a third party), $R(a_i)$ indicates the relative weight of a_i's life, and $Pr(Death(c_i, a_i))$ indicates the probability that choice c_i causes the death of a_i.

Under these conditions, in the graph presented in Fig.3, each accident-case is represented as a point inside the triangle (0,0)(1,0)(0,1), where the x-axis indicates the expected disutility for the passenger and the y-axis indicates the expected disutility for the third party.

For example, in the graph in Figure 3, point B(0.3,0.36) represents the accident case where the expected disutility of swerving is 0.3 and the expected disutility (for the third party) of proceeding on course is 0.36.

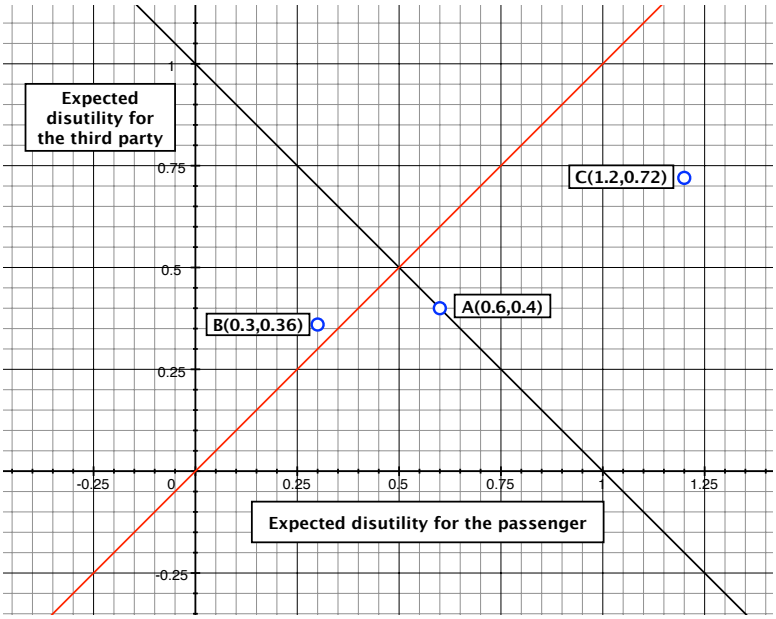

Figure 3: The AV's decision.

The line in red intercepting the origin (0,0) and the function (2) at point (0.5,0.5) represents the threshold between (a) cases in which on balance it is preferable to put the third party's life at risk, to keep the passenger safe, and (b) cases in which on balance it is preferable to put the passenger's life at risk, to keep the third party safe.

In the cases below the red line, the expected disutility of the behaviour putting the passenger's life at risk is greater than the expected disutility of the behaviour putting the third party at risk. In the cases above the line, the expected disutility of the behaviour putting the third party's life at risk is greater than the expected disutility of the behaviour putting the passenger at risk.

Note that in our scenario we have assumed that choices are between the life of a single passenger and that of a single third party. Our model can be easily extended to cover cases where more than two lives are at stake.

In these cases, the total disutility $TDis(c_i, a_i)$ could be calculated by multiplying the disutility $Dis(c_i, a_i)$ suffered as consequence of choice c_i by one agent of type a_i—where a_i is the single type of agent affected by the choice c_i—for the number $n(a_i)$ of agents of type a_i:

$$TDis(c_i, a_i) = Dis(c_i, a_i) * n(a_i) \tag{3}$$

For example, assume that there are 4 passengers in the AV, and 2 third parties, and that the expected disutilities, calculated according to formula 2,

are 0.30 for passengers and 0.36 for third parties. Using the formula (3), the total disutility for passengers ($TDis(c_1, a_1)$) and third parties ($TDis(c_2, a_2)$) will be calculated as follows:

$$TDis(c_1, a_1) = 0.30 * 4 = 1.2$$
$$TDis(c_2, a_2) = 0.36 * 2 = 0.72$$

Therefore, in this revised scenario, the expected disutilities of swerving (1.2) and holding a straight course (0.72), will be represented by point C(1.2,0.72) in Figure 3.

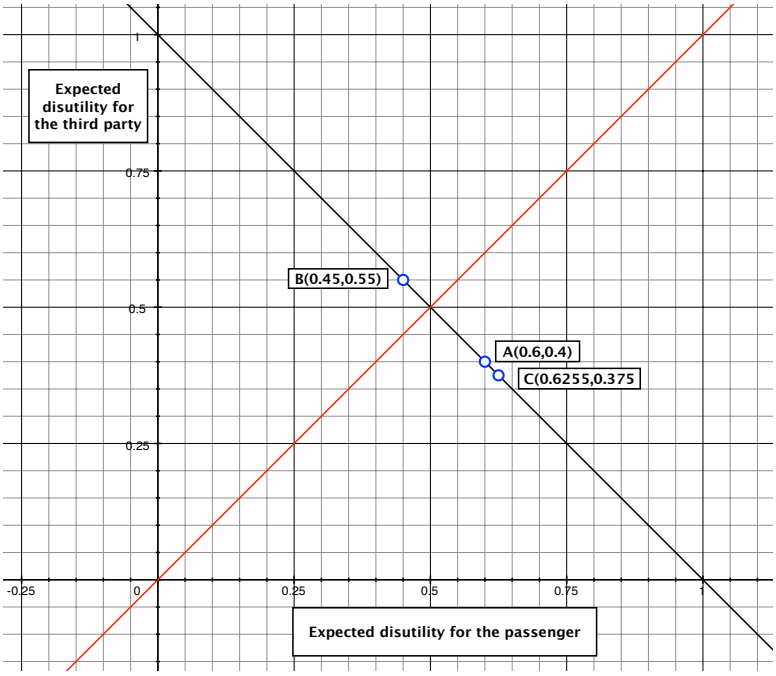

Figure 4: The AV's decision - normalized graph.

In order to obtain the normalized total disutility $NTDis(c_i, a_i)$ of the effect of choice (c_i) on agents of type a_i, we have to divide the total disutility of choice c_i (where i can be 1 or 2) by the sum of the disutilities of the two alternative choices c_1 and c_2 affecting a_1 and a_2, respectively. Thus we obtain the following formula:

$$NTDis(c_i, a_i) = \frac{TDis(c_i, a_i)}{TDis(c_1, a_1) + TDis(c_2, a_2)} \tag{4}$$

Applying formula 4 we obtain the following normalized values:

$$NTDis(c_1, a_1) = \frac{1.2}{1.2 + 0.72} = 0.625$$

$$NTDis(c_2, a_2) = \frac{0.72}{1.2 + 0.72} = 0.375$$

The values for point B can be normalized in same way[1]. The normalized value for points B(0.45,0.55) and C(0.625,0.375) are shown in Figure 4.

The possibility of setting the knob in a continuous way opens the possibility that the importance that the passenger attributes to its own life is so high that even a small risk of death for the passenger determines a disutility that is higher than the disutility corresponding to a very high risk for the pedestrian. Assume for instance that the passenger values his life 0.95 and correspondingly values 0.05 the pedestrian's life. Given this setting, in a situation in which there is a 100% probability of killing the pedestrian by proceeding and 10% probability of killing the passenger by swerving, the AV would choose to proceed. We may doubt that such a behaviour would be legally acceptable. It would be up to every legal system to determine the threshold for acceptable selfishness. Should the Ethical Knob allow for an illegal choice, not only the passenger would be liable, but also the programmer/manufacturer who has designed the knob in such a way as to allow for such a choice, as argued by Lin (2014).

3.1 A Utiliarian or a Rawlsian approach?

According to a utilitarian approach the AV should choose the course of action resulting in a lower overall disutility, i.e., a lower sum of the total expected disutilities. Thus, in order to obtain the overall disutility $ODis(c_i)$ of choice c_i, we have to sum the disutilities resulting for the choice c_i with regard to all agents $a_1, \ldots, a_n$, according to the following formula:

$$ODis(c_i) = Dis(c_i, a_1) + \ldots + Dis(c_i, a_n) \tag{5}$$

For instance, let us consider a specific scenario where 1 pedestrian and three passersby are involved. For simplicity's sake we assume that an impartial attitude is adopted, i.e., the loss of life of each of pedestrian and passerby is given the same importance, denoted by l, i.e., l is the disutility resulting from the loss of a life. Assume also that the choice of proceeding would entail with the death of the pedestrian with probability 0.9, while the choice of swerving will determine the death of each passersby with probability 0.6.

In this scenario, the overall disutility for c_1 (proceeding) will be $09 * 1l = 0.9l$, while the overall disutility of swerving would be c_2 (proceeding) will be $0.6 * 3l = 1.8l$. Consequently, the AV should choose to proceed, in order to minimize the overall disutility (the expected lives lost).

The utilitarian approach is not the only possible criterium that can be adopted. For instance, Leben (2017) advocates an algorithm inspired by Rawls's Difference Principle (1999): AV should not adopt the choice that minimises total

[1] $TDis(c_1, a_1) = 0.36 * 1 = 0.36 TDis(c_2, a_2) = 0.3 * 1 = 0.3$
$NTDis(c_1, a_1) = \frac{0.36}{0.36+0.3} = 0.55$
$NTDis(c_2, a_2) = \frac{0.3}{0.36+0.3} = 0.45$

disutility, but rather the choice that minimises the loss of the most disfavoured individual.

Following this approach we have to determine, for each choice c_i in the available choice set C, the expected disutility of the individual which is most disadvantaged by it —the one receiving the highest disutility from c_i —, which we denote as md_{c_i}. The preferred choice c_* will be the one where its most disadvantaged individual, i.e. md_{c_*}, has a better deal (an inferior disutility) than the most disadvantaged agent in any other choices: $md_{c_*} < md_{c_i}$, for all $c_i \in C$[2]

In the above scenario, choices c_1 (proceeding) and c_2 (swerving) result in the following disutilities for the agents a_1 (crossing pedestrian) and a_2, a_3, a_4 (the three passersby):

$$\begin{array}{l} Dis(c_1, a_1) = 0.9l, Dis(c_1, a_2) = 0.0l, Dis(c_1, a_3) = 0.0l, Dis(c_1, a_4) = 0.0l \\ Dis(c_2, a_1) = 0.0l, Dis(c_2, a_2) = 0.6l, Dis(c_2, a_3) = 0.6l, Dis(c_2, a_4) = 0.6l \end{array} \quad (6)$$

In choice c_1 (proceeding) the highest disutility suffered by an agent is $0.9l$ (by a_1), while in choice c_2 (swerving) the highest disutility (proceeding) by an agent is $0.6l$ (by each of a_2, a_3, a_4. Thus the Rawlsian AV should choose to swerve, rather than to proceed (as the utilitarian AV).

It seems to us that this outcome is quite counterintuitive. In fact, under the veil or ignorance —without knowing whether he or she should be the pedestrian or one of the passersby— a rational agent should not prefer to have $1.8/4 = 0.5$ chances of dying rather than only $0.9/4 = 0.225$. The puzzle results from the fact that to determine how much each agent is affected by a choice, we should not consider how many chances of dying the agent has, but rather whether the agent will die as a consequence of that choice (dying after having had a 0.9 risk of dying, or after having had a 0.6 risk does not make much difference from the concerned individual). This conclusion is strengthened if we assume that each agent has equal chances of being a pedestrian or a passerby, in different occasions.

The Rawlsian approach could appear more acceptable if we assume that the disutilities being considered represent personal injuries having the same probability, but different gravity.

For example, assume that $0,6l$ is the quantification of the damage from paraplegia (the loss or the use of both legs) and that $0.3l$ corresponds to the loss of the use of one leg. Assume that by proceeding the AV would cause with certainty the pedestrian to become paraplegic, while by swerving it would cause with certainty each one of three passersby to lose one leg each. Then it might be argued —though this conclusion is very debatable— that swerving is preferable to proceeding on equitative/equalitarian grounds.

[2]For simplicity we do not consider cases when more than one choice have the same disutility for their most disadvantaged agent, such cases have to be addressed according to the principles of lexicographic order, i.e, by considering the second most disadvantaged agent, and so on.

4 Conclusions

The moral dilemmas where AVs must choose the lesser of two evils raise many ethical and legal issues. We have explored the legal implications of letting the choice to the user, by providing AVs with an Ethical Knob.

The Ethical Knob would allow the passenger to customise the AV, by choosing between different settings corresponding to different moral approaches. The AV would correspondingly be entrusted with implementing the user's choices, while the manufacturer/programmer would enable the different settings and ensure their implementation into the AV, according to the user's choice. Therefore, with the Ethical Knob, the AV's decisions in the face of moral-legal dilemmas would depend on the customisation chosen by the user.

In comparison to the opportunity of purchasing AVs with different preprogrammed and fixed ethical settings, an Ethical Knob would provide the user with a larger set of choices that can vary over time, depending on age and number of passengers, life expectation and other factors.

Our legal analysis has shown that, with regard to their legal regime, ethically customisable AV in life and death dilemmas would significantly differ from both preprogrammed AV and humanly driven cars.

With the Ethical Knob, responsibility for fundamental ethical choices determining the functioning of the AV would be allocated to users, rather than to manufacturers. Consequently, the state-of-necessity defence would work in some cases as it would for drivers in traditional cars.

However, the fact that the settings on the knob have to be selected in advance to the accident, can affect some contexts, particularly where the distinction between action and omission could justify a non-consequentialist approach for a human driver. In this regard, ethically customisable AV would be treated similarly to preprogrammed AVs.

Regarding AVs equipped with the Ethical Knob, in principle no obligations or liabilities other than those provided for manned vehicles would fall to the producer/programmer (except in the case considered above in 3, where the programmer would be liable for designing a system enabling the user to select settings resulting in an illegal behaviour by the AV.

This could facilitate the placement of AVs on the market. Furthermore, the Ethical Knob may improve users' acceptance of AVs, giving users the ability to choose the moral algorithm that reflects their moral attitudes and convictions.

Some interesting issues emerge regarding the possibility of choosing settings such that the importance of a passenger's life is higher than that of a third party.

In the first place, this should have an impact on insurance, since these settings would increase the risk of injuries to third parties (moral hazard). Suppose, for example, that a person turns the knob to an extreme egoistic mode, as by valuing his or her life 99 times more than the life of a third party. Suppose also that in particular circumstances swerving can lead to the death of the passenger only with a 2% probability, but holding course would certainly lead to the death of a third party. Under such circumstances the AV would choose to hold

its course. Therefore, the third party's life would be sacrificed under conditions in which a different choice would most probably not have caused any death (since there is a 98% chance that swerving would not affect the passenger's life). Since setting the knob to the passenger's benefit increases the chances of accidents, this should at least lead to a higher insurance premium (or a more limited coverage) for those who preselect that preference.

We may also wonder to what extent in such a scenario, the state-of-necessity defence could still apply, given that the risk of the passenger losing his or her life is so remote (by comparison with the certain sacrifice of the third party's life). Regulation or case law should determine what level of user-selected egoism could lead to an AV behaviour that could expose the user to criminal or civil liability.

The authors would like to thank the anonymous reviewers for their many valuable comments and suggestions.

References

Bonnefon JF, Shariff A, Rahwan I (2015) Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? arXiv preprint arXiv:151003346

Bonnefon JF, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352(6293):1573–1576

Foot P (1967) The problem of abortion and the doctrine of double effect. Oxford Review 1(5)

Gogoll J, Müller JF (2016) Autonomous cars: In favor of a mandatory ethics setting. Science and Engineering Ethics DOI 10.1007/s11948-016-9806-x, URL http://dx.doi.org/10.1007/s11948-016-9806-x

Goodall NJ (2016) Away from trolley problems and toward risk management. Applied Artificial Intelligence 30(8):810–821

Leben D (2017) A rawlsian algorithm for autonomous vehicles. Ethics and Information Technology

Lin P (2014) Here's a terrible idea: robot cars with adjustable ethics settings. Wired com Available via http://www wired com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings

Lin P (2016) Why ethics matters for autonomous cars. In: Autonomous Driving, Springer, pp 69–85

Millar J (2015) Technology as moral proxy: Autonomy and paternalism by design. IEEE Technology and Society Magazine 34(2):47–55

Nyholm S, Smids J (2016) The ethics of accident-algorithms for self-driving cars: an applied trolley problem? Ethical Theory and Moral Practice pp 1–15, DOI 10.1007/s10677-016-9745-2, URL http://dx.doi.org/10.1007/s10677-016-9745-2

Rawls J (1999) A Theory of Justice, revised edn. Oxford University Press

Santoni de Sio F (2017) Killing by autonomous vehicles and the legal doctrine of necessity. Ethical Theory and Moral Practice 20(2):411–429, DOI 10.1007/s10677-017-9780-7, URL http://dx.doi.org/10.1007/s10677-017-9780-7

Thomson JJ (1976) Killing, letting die, and the trolley problem. The Monist 59(2):204–217