No Ethics Settings for Autonomous Vehicles

Bracanović, Tomislav

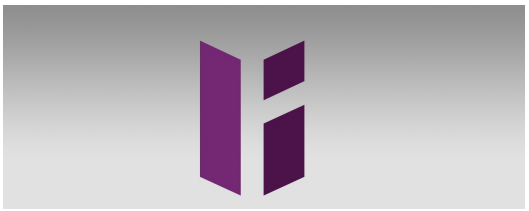
Source / Izvornik: **Magyar Filozófiai Szemle / Hungarian Philosophical Review, 2019, 63, 47 - 60**

Journal article, Published version
Rad u časopisu, Objavljena verzija rada (izdavačev PDF)

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:261:849567

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2025-02-22**



Repository / Repozitorij:

Repository of the Institute of Philosophy

Tomislav Bracanović

No Ethics Settings for Autonomous Vehicles*

Abstract

Autonomous vehicles (AVs) are expected to improve road traffic safety and save human lives. It is also expected that some AVs will encounter so-called dilemmatic situations, like choosing between saving two passengers by sacrificing one pedestrian or choosing between saving three pedestrians by sacrificing one passenger. These expectations fuel the extensive debate over the ethics settings of AVs: the way AVs should be programmed to act in dilemmatic situations and who should decide about the nature of this programming in the first place. In the article, the ethics settings problem is analyzed as a trilemma between AVs with personal ethics setting (PES), AVs with mandatory ethics setting (MES) and AVs with no ethics settings (NES). It is argued that both PES and MES, by being programmed to choose one human life over the other, are bound to cause serious moral damage resulting from the violation of several principles central to deontology and utilitarianism. NES is defended as the only plausible solution to this trilemma, that is, as the solution that sufficiently minimizes the number of traffic fatalities without causing any comparable moral damage.

Keywords: autonomous vehicles, ethics settings, utilitarianism, deontology, moral damage

I. INTRODUCTION

Autonomous vehicles (AVs) are expected to improve road traffic safety and reduce the number of traffic fatalities, especially those caused by human factors such as alcohol or drugs abuse, carelessness, fatigue and poor driving skills. It is also expected that some AVs – despite their enhanced reliability made possible by AI algorithms, interconnectedness, sophisticated sensors and similar tech-

* The first version of this article was presented at the *Zagreb Applied Ethics Conference*, organized in June 2019 by the Society for the Advancement of Philosophy and the Institute of Philosophy in Zagreb. I am grateful to members of the audience for their comments. I am also grateful to an anonymous referee of the *Hungarian Philosophical Review* for useful suggestions.

nologies – are bound to encounter dilemmatic situations of having to choose the lesser of two (or more) evils. To mention some standard hypothetical examples: an AV might have to decide whether to sacrifice one pedestrian to save three others, to save two pedestrians by sacrificing the passenger of the vehicle or to sacrifice an elderly person to save a child. Hypothetical examples like these, usually formulated in terms of the classic trolley problem (Foot 1967, Thomson 1976), find themselves at the center of the debate over ethics settings of AVs: How should AVs be programmed to react in dilemmatic situations and who should decide about the nature of this programming in the first place? Many scholars believe that this debate (useful reviews are Millar 2017 and Nyholm 2018a, 2018b) is of great practical significance. According to Awad and colleagues:

> Never in the history of humanity have we allowed a machine to autonomously decide who should live and who should die, in a fraction of a second, without real-time supervision. We are going to cross that bridge any time now, and it will not happen in a distant theatre of military operations; it will happen in that most mundane aspect of our lives, everyday transportation. Before we allow our cars to make ethical decisions, we need to have a global conversation to express our preferences to the companies that will design moral algorithms, and to the policymakers that will regulate them. (Awad et al. 2018. 63)

It is argued in the present article that introduction of AVs with any type of ethics settings that would enable them to "decide who should live and who should die" (Awad et al. 2018. 63) is bound to cause serious moral damage, construed here as a violation of several principles central to both the deontological and utilitarian ethical traditions. A similar argument can be found in the report on *Automated and Connected Driving*, published by the Ethics Commission appointed by the German Federal Minister of Transport and Digital Infrastructure (BMVI 2017). The report emphasizes that "human lives must not be 'offset' against each other" and finds it impermissible "to sacrifice one person in order to save several others" (BMVI 2017. 18). The difference between the present article and the German report, however, lies in their respective premises: whereas the premises of the German report are predominantly deontological, this article's premises are deontological *and* utilitarian. The article, in other words, elaborates upon the deontological case from the German report, but it also develops an additional utilitarian case. The primary purpose of the article, however, is not to decide which ethical position, deontology or utilitarianism, is more promising when it comes to rebutting the idea of AVs with ethics settings. Rather, its primary purpose is to explicate the range and diversity of arguments against ethics settings and to suggest that – despite all the "global conversation" (Awad et al. 2018. 63) and philosophical efforts – AVs with ethics settings will remain not

only a bridge that we should not cross, but most likely a bridge that most people will never seriously intend to cross.

The present article consists of six sections. Following this section, section II describes the problem as a trilemma between three types of ethics settings: personal ethics setting (PES), mandatory ethics setting (MES) and no ethics settings (NES). Section III develops deontological and utilitarian arguments against PES and section IV does the same with respect to MES. In section V, NES is defended as the only plausible solution to this trilemma. Section VI concludes the article by summarizing the main points.

II. THE TRILEMMA

Consider the trilemma between three types of ethics settings (the abbreviations PES and MES, with slight modifications of what they refer to, are borrowed from Gogoll and Müller 2017):

PES Personal ethics setting. Ethics settings should be chosen individually by the AV's passengers. Although "personal" is not by definition "egoistic" or "selfish", it is assumed here that PES is predominantly selfish, that is, it is programmed to save the passengers of the AV even at the expense of sacrificing a greater number of other people.

MES Mandatory ethics setting. Ethics settings for all AVs should be the same and chosen and enforced by the state. It is assumed here that MES impartially distributes harms and benefits among all those affected by its decisions. For example, it always saves the greatest number of lives, even at the expense of sacrificing the passengers of the AV.

NES No ethics settings. AVs should have no ethics settings, in the sense that they should have no pre-programmed rules enabling them to choose one human life over the other.

III. THE CASE AGAINST PES

Despite its coherence with individual autonomy as one of the most fundamental deontological principles, deontologists would reject PES as long as its decision-making were guided by the selfish interests of the AV's passengers. From the deontological point of view, acting with selfish motives is the antithesis of moral behavior. That AVs with PES would in most cases exemplify this antithesis is not just armchair speculation about human nature but something corroborated by empirical research. For example, in one poll, 64% of participants answered that they would even sacrifice a child in order to save themselves (Millar

2017. 25); other studies reveal that "[a]lthough people tend to agree that everyone would be better off if AVs were utilitarian (in the sense of minimizing the number of casualties on the road), these same people have a personal incentive to ride in AVs that will protect them at all costs" (Bonnefon et al. 2016. 1575). As a matter of fact, in order for it to fail by deontological standards, especially those set by Immanuel Kant (1785/1996), an AV with PES need not be *sensu stricto* selfish, that is, contributing exclusively to the well-being of its passenger. Just as unacceptable would be any other arbitrary or heteronomous motivation or reason – for example, positive or negative attitudes towards someone's race, sex, ethnicity or age – for distinguishing between traffic participants whose lives are worth saving from those whose lives are not worth saving.

PES also violates another important deontological principle: the prohibition against using persons "merely as means" (sometimes referred to as the "personhood" principle). In Kant's words, a human being "can never be used merely as a means by anyone (not even by God) without being at the same time himself an end" (1788/1996. 245). If I program my AV to systematically sacrifice anyone else in order to save my own life, this obviously amounts to using other persons merely as means. People treating each other as means, of course, is morally unproblematic as long as they do not treat each other *merely* as means, in the sense that everyone involved either explicitly agrees to a specific scheme of (inter) action or that their consent can be reasonably presumed (O'Neill 1994. 44). For example, I use the delivery driver as a means to get my pizza and he uses me as a means to earn his wages. The problem appears when people are treated *merely* as means and would not consent to such treatment if they were asked. For example: A and B survived a plane crash on a desert island. A kills B in his sleep, so he can eat him and survive until the rescuers arrive. B did not consent – and probably would not if A asked him – to be used in this way. PES is structurally similar and, for this reason, similarly problematic. One cannot reasonably presume that any person – in the other vehicle, or on the sidewalk or crosswalk – has consented to be killed (to be used merely as a means), so that I can continue living. I can reasonably presume my delivery driver's consent to be used as a means to get my pizza, but I cannot presume his consent to be run over by my AV to stop it from crashing into the back of a truck.

Partiality is one of the clearest utilitarian deficits of PES. Utilitarians insist on "the greatest happiness for the greatest number", but they also insist that this happiness is achieved in an impartial way. In John Stuart Mill's formulation: "[T]he happiness which forms the utilitarian standard of what is right in conduct is not the agent's own happiness, but that of all concerned" and "between his own happiness and that of others, utilitarianism requires him to be as strictly impartial as a disinterested and benevolent spectator" (1863/1998. 64). Peter Singer uses the "scales" metaphor: "True scales favour the side where the interest is stronger or where several interests combine to outweigh a smaller

number of similar interests, but they take no account of whose interests they are weighing" (2011. 20–21). An AV with PES that prioritizes its passengers' lives and interests over all other lives and interests – an option, as we have seen, that would be adopted by the majority of AV passengers – would obviously violate this utilitarian requirement of strict impartiality and disinterested benevolence.

A more serious utilitarian deficit of PES is its strong tendency – in comparison to other types of ethics settings – to bring about the worst possible consequences. If most AVs are set to protect their passengers' lives at all costs, including the cost of sacrificing any number of other lives, that should unquestionably, in the long run, increase the total number of traffic fatalities. This outcome is diametrically opposed to the fundamental utilitarian (consequentialist) principle of minimizing suffering and maximizing happiness for the greatest number of people possible. An argument to the same effect, presented in game-theoretical terms, is offered by Gogoll and Müller (2017). They maintain that allowing people to personally choose their own ethics settings would create "prisoner dilemma" circumstances in which everyone's probability of dying in traffic increases. Their basic point is this: even individuals disposed to choose "moral" PES (sacrificing themselves to save the greater number of others), as opposed to "selfish" PES (sacrificing any number of others to save themselves), would at some point realize that they are taken advantage of by selfish individuals. In this kind of environment, guided by rationality and in pursuit of their own interest, they would eventually switch to "selfish" ethics settings themselves, contributing thus to the creation of "a world in which nobody is ready to sacrifice themselves for the greater number" and "the number of actual traffic casualties is necessarily higher" (Gogoll–Müller 2017. 694). The proposed solution to this dilemma – to be analyzed in the next section – is MES:

> This leaves us with the classical solution to collective action problems: governmental intervention. The only way to achieve the moral equilibrium is state regulation. In particular, the government would need to prescribe a mandatory ethics setting (MES) for automated cars. The easiest way to implement a MES that maximizes traffic safety would be to introduce a new industry standard for automated cars that binds manufacturers directly. The normative content of the MES, that we arrived at through a contractarian thought experiment, can easily be summarized in one maxim: *Minimize the harm for all people affected!* (Gogoll–Müller 2017. 695)

IV. THE CASE AGAINST MES

The deontological deficits of MES are practically the mirror image of the deontological deficits of PES: whereas the major problem with PES is not autonomy but selfishness, the major problem of MES is not selfishness but autonomy.

As the German report on *Automated and Connected Driving* correctly recognizes, MES implies that "humans would, in existential life-or-death situations, no longer be autonomous but heteronomous" and that the state would act "in a very paternalistic manner and prescribing a 'correct' ethical course of action" (BMVI 2017. 16). MES would basically suspend an individual's capacity for ethical decision-making in situations – those with human lives at stake – in which the exercise of this capacity might be most needed. In other words, autonomous decision-making and moral agency would be substituted by algorithmic ("heteronomous") decision-making and preprogrammed agency. Since the specifics of this decision-making, by the definition of MES, would be prescribed and enforced by the state, it may actually be inadequate to talk about it as *moral* or *ethical* decision-making – in the same way as it would be erroneous to talk about any state prescribed and enforced norms as moral or ethical. In short, deontologists could claim that MES, as a consequence of its suspension of individual autonomy and moral agency, is actually a negation of ethics and should not be classified as an "*ethics* setting" at all.

An equally important deontological deficit of MES is that it implies using persons merely as means, in the sense of sanctioning a practice of sacrificing some persons – when traffic circumstances dictate it – to save the greater number of others. The fact that this would not be done by other persons (as was the case with PES), but by the state, is morally irrelevant. If a human being, as Kant said, "can never be used merely as a means by anyone (not even by God)", then they cannot be used merely as a means even by the state. The German report similarly points out that "offsetting of victims" by AVs is impermissible because "the individual is to be regarded as 'sacrosanct'" and "equipped with special dignity" (BMVI 2017. 18–19). It is important to notice that the wrongness of using persons merely as means here does not essentially stem from the fact that it would be performed by machines (which is a common ethical objection to many similar uses of AI systems). It would be wrong even if it was performed by human beings. Imagine that a time machine is invented that allows humans, at any given moment, to "freeze" time and everything that happens. They can "freeze" dilemmatic situations with AVs before they play out and allow human experts – some kind of a time travelling ethics committee – enough time to decide how to resolve them (for example, whether to sacrifice pedestrians or passengers). Assuming that persons affected by these decisions would not be consulted, the time travelling ethics committee would be treating them merely as means in the same way that MES would.

A possible reply to "autonomy" and "personhood" objections is that their force diminishes if all or the majority of citizens decide, through some kind of democratic procedure, that they wish to trade parts of each individual's autonomy and personhood for the reduction of everyone's chances of being killed in traffic. The problem with this reply is well-known from ethical debates on a

variety of sensitive issues like abortion, euthanasia or capital punishment: the majority opinion is not necessarily the morally right opinion. A public referendum with any percentage of votes – tight votes especially – either approving or disapproving any of these practices does not settle the fundamental ethical question of their rightness or wrongness (except, maybe, for radical ethical relativists). As an institutional arrangement that will require almost daily choices between human lives, MES would surely become an extremely sensitive issue likely to split public opinion. However, in view of the diversity and value pluralism of contemporary democratic societies, it seems unsatisfactory to use any form of democratic decision-making as a tiebreaker for moral disputes with far-reaching consequences like the one over MES. For the same reason, is does not seem promising to use it to neutralize deontological objections as complex as autonomy or personhood.

The main problem with MES, from the deontological perspective, is the fact that it is a utilitarian scheme of action and all such schemes, in John Rawls's formulation, have to be rejected because they disregard "the distinction between persons" (1971/1999. 24). According to Rawls, it is impermissible "that the sacrifices imposed on a few are outweighed by the larger sum of advantages enjoyed by many" (1971/1999. 3) and, "under most conditions, at least in a reasonably advanced stage of civilization, the greatest sum of advantages is not attained in this way" (1971/1999. 23). Nevertheless, it might be too hasty to conclude that MES, despite its central goal of minimizing the harm for all people affected, would be mechanically taken on board by utilitarians. The way in which MES would accomplish this goal is likely to have harmful side effects that most utilitarians tend to invoke when they dismiss some other, in many respects similar, proposals. As an initial illustration, consider the following hypothetical example:

> You have five patients in the hospital who are dying, each in need of a separate organ. […] You can save all five if you take a single healthy person and remove his heart, lungs, kidneys, and so forth, to distribute to these five patients. Just such a healthy person is in room 306. He is in the hospital for routine tests. Having seen his test results, you know that he is perfectly healthy and of the right tissue compatibility. […] The other five patients can be saved only if the person in Room 306 is cut up and his organs distributed. In that case, there would be one dead but five saved. (Harman 1977. 3–4)

In terms of the number of lives to be saved, cutting up the person in room 306 seems to make perfect utilitarian sense: "one dead but five saved" sounds much better than "one saved but five dead". Most utilitarians, however, are more refined than that and numbers are not the only thing that matters in their moral reasoning. They tend to dismiss proposals like cutting up the person in room 306, because they believe that any similar practice, once it is allowed and be-

comes publicly known, is likely to have a series of harmful side effects. According to rule utilitarians like Richard Brandt (1965/2003), for example, a rule that allows one healthy person to be sacrificed in order to save five dying patients might, in the long run, bring about an even greater loss of lives (e.g. due to the growing distrust of doctors or the fear of visiting hospitals). An advocate of R. M. Hare's (1981) two-level utilitarianism could claim that doctors, due to their inherent human limitations and biases, would more often than not make wrong judgments about exactly who and when should be sacrificed, so that the greatest number of lives can be saved. Since they are very likely to have catastrophic consequences, calculations like these should not be allowed to be part of doctors' everyday work. Peter Singer's preference utilitarianism also leaves plenty of room for rejection of similar proposals and practices:

> If […] we decided to perform extremely painful or lethal scientific experiments on normal adult humans, kidnapped at random from public parks for this purpose, adults who entered parks would become fearful that they would be kidnapped. The resultant terror would be a form of suffering additional to the pain of the experiment. (Singer 2011. 51–52)

> If I am a person, I know that I have a future. I also know that my future existence could be cut short. If I think that this is likely to happen at any moment, my present existence will be fraught with anxiety and will presumably be less enjoyable than if I do not think I am likely to die for some time. If I know that people like myself are very rarely killed, I will worry less than if the opposite is the case. (Singer 2011. 77)

The utilitarian logic behind the examples mentioned so far can be captured as follows: Although an action may have some positive immediate effects (for example, five lives saved at the expense of one), there is an overriding reason against performing that action as long as it, once becoming publicly known, is likely to have negative side effects across the population at large and continuing indefinitely into the future (for example, the resultant terror, fear and anxiety at the individual and social level). Another useful illustration of this logic is the hypothetical example of "survival lottery" by John Harris (1975/1986). We are invited to imagine two dying patients, Y and Z, trying to persuade doctors to save their lives by acquiring healthy organs in a unique way:

> Y and Z put forward the following scheme: they propose that everyone be given a sort of lottery number. Whenever doctors have two or more dying patients who could be saved by transplants, and no suitable organs have come to hand through "natural" deaths, they can ask a central computer to supply a suitable donor. The computer will then pick the number of a suitable donor at random and he will be killed so that the lives of two or more others may be saved. (Harris 1975/1986. 89)

One possible reason for rejecting "the institution of the survival lottery", according to Harris (1975/1986. 92), is that its "harmful side effects in terms of terror and distress to victims, witnesses, and society generally" would be similar to the harmful side effects "occasioned by doctors simply snatching passers-by off the streets and disorganizing them for the benefit of the unfortunate." This "lottery scheme", as Harris emphasizes, "would eliminate the arbitrariness of leaving the life and death decisions to the doctors, and remove the possibility of such terrible power falling into the hands of any individuals, but the terror and distress would remain" (1975/1986. 92). In what follows, it will be argued that MES bears sufficient resemblance to actions like "cutting up the person in room 306", "kidnaping people at random from public parks for lethal experiments" and the "survival lottery" itself, to be rejected on the very same utilitarian grounds.

MES and "cutting up the person in room 306" are analogous due to the decisive role that randomness plays in them. Assume that the person in room 306 ends up being cut up and his organs distributed to five dying patients. It happened only because he, accidentally, visited a particular hospital on a particular day and was outnumbered by five dying patients that were, accidentally, in the same place at the same time and could be saved by his organs. Had he decided to visit another hospital (or the same hospital on another day), he would still be alive. MES would also sacrifice a person only because she, accidentally, crossed a particular street at a particular time and was outnumbered by several other people that were, purely by chance, in the same place at the same time and could be saved by sacrificing her. Had she decided to cross some other street (or the same street at a different time), she would still be alive. This kind of accidental factor or randomness, if allowed (and publicly announced) to influence life or death decisions in everyday circumstances, would undoubtedly cause enough "terror and distress to victims, witnesses, and society generally" to justify the utilitarian rejection of any similar scheme of action.

There is something more problematic with MES than with "cutting up the person in room 306". The conditions that have to be met, namely, for doctors to even begin considering the proposal of "cutting someone up" would be exceptional: What is the probability of (a) a healthy person (b) visiting the hospital for routine tests, (c) having his tissue compatible with five patients (d) each in need of a different organ, that are (e) already present in the hospital? This probability must be extremely low, but there is no doubt that most utilitarians would still reject any similar scheme of action – especially if it should become publicly known – as not worth the risk of harmful side effects. The problem with MES is that the probability of finding oneself in a dilemmatic situation, potentially as the person that has to be sacrificed by an AV, will be significantly higher. This much should be clear already from the fact that a large portion of the population participates in or is somehow affected by road traffic on a daily basis. Moreover, the probability of such an event will become even higher if AVs, as the Institute of

Electrical and Electronics Engineers (IEEE 2012) has predicted, "will account for up to 75 percent of cars on the road by the year 2040." What follows is that with state-wide implementation of MES, very few will be able to say that people like themselves "are very rarely killed", that their "future existence" is unlikely to be "cut short" and that they, therefore, have no reason to worry about MES.

As the final variation of the same utilitarian argument against MES, imagine MES 2.0 – an advanced version of MES that takes into account not only the number of people involved in dilemmatic situations (either as passengers or as pedestrians), but additional factors as well, such as their health status, age, profession, number of children and criminal record. The collection and use of such sensitive data – essentially in order to profile individuals for their suitability to be saved or sacrificed for the greater good – would be perceived by the general public as something negative and intimidating. Moreover, if its functioning will depend on technologies like machine learning or self-learning algorithms, it could be extremely difficult, from the technical point of view, to explain to the public how and on the basis of which data MES 2.0 makes its life or death decisions. A purely technical issue like this – also known as the "black box" problem of algorithms – would easily morph into a moral and political issue: any non-transparency, inexplicability or secrecy related to tools like MES, especially when they are controlled by state officials, tends to fuel suspicions and fear of things like corruption, discrimination or even totalitarianism. Bearing in mind, moreover, that AVs are "the first robots to be integrated with society at any significant scale" that might "set the tone for other social robotics, especially if things go wrong" (Lin–Jenkins–Abney 2017. ix), these suspicions and fear provide a solid utilitarian argument against MES.

It is possible to remain sceptical about the idea of MES as a cause of distress, anxiety and fear. Moreover, the opposite claim could be argued for: given that accidents that already happen with conventional vehicles do not trigger any systematic distress, anxiety and fear, AVs with MES could actually, by minimizing everyone's chances of being killed in traffic, prevent the occurrence of any similar distress, anxiety and fear. One problem with such a defense of MES is that practices like "cutting up the person in room 306" or the "survival lottery" could be justified in a similar way (by arguing, for example, that they would improve the chances of survival of all hospitalized persons or all members of society), but they would still be perceived as serious and morally unacceptable sources of distress, anxiety and fear. Another problem is that individuals might not care (although perhaps irrationally) about the statistical advantages of MES as much as they care about some other things it might interfere with, like the freedom to make their own decisions in life or death situations, a desire to protect their own lives or the lives of their family members first, or even – as we shall see in the next section – a commitment to certain moral principles and values. It should be emphasized, however, that the objective of this section was not to answer

the empirical question about the psychological effects of MES. Its objective was primarily conceptual: to identify similarities between the idea of MES and hypothetical scenarios that utilitarians themselves tend to reject and to show, in this way, that the moral damage potentially generated by MES need not be only deontological, but also utilitarian.

V. THE CASE FOR NES

Implementing either PES or MES, due to their unavoidable violation of several principles central to both deontology and utilitarianism, is bound to cause serious moral damage. In a nutshell, the deontological deficits of PES are the expected selfishness and using other persons merely as means, while its utilitarian deficits are the expected partiality and the tendency to bring about the worst possible outcomes in terms of the number of traffic fatalities. The deontological deficits of MES are the suspension of individual autonomy and using other persons (this time by the state) merely as means, while its crucial utilitarian deficit is the high potential to bring about harmful side effects – like distress, anxiety and fear at individual and social level – which most utilitarians anticipate and invoke when they reject some highly similar schemes of action. The presence of moral damage constituted by these moral deficits solves our initial trilemma: PES and MES have to be excluded and NES – that is, AVs unable to choose one human life over the other – remains the only plausible option.

It should be recognized that MES, thanks to its impartial distribution of harms and benefits among all those affected by its decisions, outcompetes both PES and NES in minimizing the number of traffic fatalities. However, a combination of two reasons, one statistical and the other one moral, is what makes NES the only plausible option. The statistical reason is that, when it comes to minimizing the number of traffic fatalities, NES outcompetes PES and is still, therefore, the second-best solution to how AVs should behave in dilemmatic situations. (Remember that PES has a practically inbuilt tendency to maximize the number of traffic fatalities whenever that saves the AV's passenger.) The moral reason should be familiar by now: NES causes no moral damage comparable to the one caused by either PES or MES. NES should be preferred to its alternatives, simply put, thanks to the best ratio of the expected success in minimizing the number of traffic fatalities to the expected range of its moral damage. In order to explicate this point further, consider the following hypothetical case by Bonnefon, Shariff and Rahwan:

> Say that two competing companies market self-driving cars that both eliminate 80% of fatalities, but one company's cars split the remaining fatalities equally between passengers and pedestrians, whereas the other company's cars split the remaining

> fatalities nine-to-one in favor of their passengers. Consumers would flock to the cars of the second company, and pedestrian risks would gradually inflate to unacceptably unfair levels. (Bonnefon at al. 2019. 504)

Under the assumption that AVs without any kind of ethics settings eliminate 80% of traffic fatalities and that AVs with some kind of ethics settings eliminate the remaining 20%, how can it be that this additional reduction of traffic fatalities does not suffice to compensate for the moral damage that any ethics settings might cause? How can saving moral principles or abstract values be more important than saving human lives? One answer to these questions could be hiding in the hypothetical case itself: If using PES to eliminate the additional 20% of traffic fatalities is considered unacceptably unfair to pedestrians as a *group*, then using MES to achieve the same 20% improvement should be considered unacceptably unfair to any *individual* (passenger or pedestrian) killed by an AV only because she happened to be (from her perspective) in the wrong place at the wrong time (although in the right place and the right time from the perspective of those saved by sacrificing her life). Illustrated by analogy with doctors cutting up one person as a "donor" and distributing his organs to five dying patients: It would surely be unfair to select this person from a specific group of potential donors (for example, already hospitalized patients, persons over 50 or people without children), but it does not seem any fairer to select this person at random from visitors of public parks, people on the street or – for that matter – the general population.

Another answer is more general: to save as many lives as possible is desirable, but the way they are saved is not morally irrelevant and it may, depending on the situation, constitute a reason against saving them. Consider negotiating with terrorists, torturing kidnappers, paying ransoms, collective punishment, wiretapping of ordinary citizens, buying and selling of newborns for adoption, etc. Although practices like these, in certain circumstances, could save lives, they tend to be widely rejected as morally unacceptable. This rejection is typically defended in either deontological or utilitarian terms, by claiming that allowing such practices violates basic human rights or that it sets dangerous precedents with harmful side effects. It is interesting, moreover, that some of these practices are considered morally unacceptable even in emergency situations like war. It is particularly interesting that there are numerous voices, among both scholars and the general public, opposed to any wartime use of military robots or autonomous weapons. One frequently mentioned reason for this opposition is that these weapons could not distinguish combatants as legitimate targets from innocent civilians as illegitimate ones. The lesson for the ethics settings debate, at the very least, is the following: if unintentionally sacrificing innocent lives is a serious reason to reject autonomous weapons in extraordinary situations such as war, it is too unrealistic to expect any serious acceptance of AVs programmed to intentionally choose one innocent human life over the other in ordinary situations like daily traffic.

VI. CONCLUSION

The primary purpose of this article was not to decide which ethical position, deontology or utilitarianism, provides a more fertile ground for building a case against AVs with ethics settings. Its primary purpose was to argue that any type of ethics settings capable of choosing one human life over the other is bound to cause serious moral damage resulting from the violation of several principles central to both deontology and utilitarianism. AVs without ethics settings are the preferred solution because that option sufficiently minimizes the number of traffic fatalities without causing any comparable moral damage. The overall conclusion of the article is that AVs with ethics settings will remain not only a bridge that we should not cross, but most likely a bridge that most people will never have a serious intention of crossing.

REFERENCES

Awad, Edmond – Sohan Dsouza – Richard Kim – Jonathan Schulz – Joseph Henrich – Azim Shariff – Jean-François Bonnefon – Iyad Rahwan 2018. The Moral Machine Experiment. *Nature*. 563. 59–64.

BMVI 2017. *Automated and Connected Driving*. Bundesministerium für Verkehr und digitale Infrastruktur.
www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile

Bonnefon, Jean-François – Azim Shariff – Iyad Rahwan 2016. The Social Dilemma of Autonomous Vehicles. *Science*. 352(6293). 1573–1576.

Bonnefon, Jean-François – Azim Shariff – Iyad Rahwan 2019. The Trolley, the Bull Bar, and Why Engineers Should Care about the Ethics of Autonomous Cars. *Proceedings of the IEEE*. 107(3). 502–504.

Brandt, Richard B. 2003 [1965]. Toward a Credible Form of Utilitarianism. In Stephen Darwall (ed.) *Consequentialism*. Oxford, Blackwell. 207–235.

Foot, Philippa 1967. The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*. 5. 5–15.

Gogoll, Jan – Julian F. Müller 2017. Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Science and Engineering Ethics*. 23(3). 681–700.

Hare, Richard M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford, Oxford University Press.

Harman, Gilbert 1977. *The Nature of Morality: An Introduction to Ethics*. New York, Oxford University Press.

Harris, John 1986 [1975]. The Survival Lottery. In Peter Singer (ed.) *Applied Ethics*. New York, Oxford University Press. 87–95.

IEEE 2012. Look Ma, No Hands! Institute of Electrical and Electronics Engineers. https://www.ieee.org/about/news/2012/5september-2-2012.html

Kant, Immanuel 1785/1996. *Groundwork of The Metaphysics of Morals*. In Immanuel Kant: *Practical Philosophy*. Transl. by Mary J. Gregor. Cambridge, Cambridge University Press.

Kant, Immanuel 1788/1996. *Critique of Practical Reason*. In Immanuel Kant: *Practical Philosophy*. Transl. by Mary J. Gregor. Cambridge, Cambridge University Press.

Lin, Patrick – Ryan Jenkins – Keith Abney 2017. Preface. In Patrick Lin – Ryan Jenkins – Keith Abney (eds.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York/NY, Oxford University Press. ix–xiii.

Mill, John Stuart 1863/1998. *Utilitarianism*. Oxford, Oxford University Press.

Millar, Jason 2017. Ethics Settings for Autonomous Vehicles. In Patrick Lin – Ryan Jenkins – Keith Abney (eds.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York, Oxford University Press. 20–34.

Nyholm, Sven 2018a. The Ethics of Crashes with Self-Driving Cars: A Roadmap, I. *Philosophy Compass*. 13(7). https://doi.org/10.1111/phc3.12507

Nyholm, Sven 2018b. The Ethics of Crashes with Self-Driving Cars: A Roadmap, II. *Philosophy Compass*. 13(7). https://doi.org/10.1111/phc3.12506

O'Neill, Onora 1994. A Simplified Account of Kant's Ethics. In James E. White (ed.) *Contemporary Moral Problems*. St. Paul, West Publishing Company. 43–48.

Rawls, John 1971/1999. *A Theory of Justice. Revised Edition*. Cambridge, Harvard University Press.

Singer, Peter 2011. *Practical Ethics*. 3rd edition. Cambridge, Cambridge University Press.

Thomson, Judith Jarvis 1976. Killing, Letting Die, and the Trolley Problem. *Monist*. 59. 204–217.