Self-Driving Vehicles, Autonomy and Justice

Tomislav Bracanović*
tbracanovic@ifzg.hr
https://orcid.org/0000-0001-8168-2194

https://doi.org/10.31192/np.22.3.6
UDK/UDC: 629.33:179
007.52:179
Izvorni znanstveni rad /
Original scientific paper
Primljeno/Received:
10. travnja 2024. / Apr 10, 2024
Prihvaćeno/Accepted:
5. lipnja 2024. / Jun 5, 2024

Self-driving vehicles, as it is widely believed, will need to be equipped with so-called ethics settings, reducing the fatalities in unavoidable crash situations. These ethics settings, as it is also widely believed, should be mandatory and impartially distribute fatalities between vehicle passengers and pedestrians, preferring the side where more lives can be saved. This paper criticizes both of these beliefs, focusing on the tension mandatory ethics settings are about to generate between values of autonomy and justice. The central argument is that mandatory ethics settings, as long as they rely only on the number of lives to be saved or sacrificed, fail to respect one elementary principle of justice (»equals should be treated equally and unequals unequally«) by disregarding some significant differences between self-driving vehicle passengers and pedestrians. After rejecting a possible reply to this argument (that mandatory ethics settings could also rely on certain qualitative features of participants in unavoidable crash situations), it is concluded that no generally acceptable normative solution to the ethics settings problem is likely and that technological approaches to reducing fatalities in traffic with self-driving vehicles should be preferred.

Key words: *autonomy, discrimination, equality, justice, self-driving vehicles.*

* Tomislav Bracanović, PhD, senior research fellow, Institute of Philosophy; Address: Ulica grada Vukovara 54, HR-10000 Zagreb, Croatia.

*Introduction***

Artificial intelligence is a challenge to many of our highly cherished values. It is recognized and addressed as such by the growing number of scholars working on the emerging field AI ethics,[1] as well as by many professional and global organizations.[2] This paper analyzes how one of the most discussed applications of artificial intelligence – self-driving vehicles with their specific ethics settings – is about to increase the tension between autonomy and justice as two central ideas of ethics and political philosophy. Section (1) sets the stage for the rest of the paper by illustrating the place of autonomy and justice in various theoretical and practical contexts. Section (2) presents the ethics settings of self-driving vehicles debate, with its central dilemma as to whom the vehicles should sacrifice (passengers or pedestrians) in unavoidable crash situations and three standard proposals for its solution: no ethics settings (NES), personal ethics settings (PES) and mandatory ethics settings (MES). Section (3) presents an argument against MES, suggesting that it, as long as it relies only on the number of people to be saved or sacrificed, fails to respect one elementary principle of justice (»equals should be treated equally and unequals unequally«) by disregarding some significant differences between the self-driving vehicle passengers and pedestrians. Section (4) considers a possible reply to this argument (that MES could be made more consistent with justice by allowing it to make its decisions based on certain qualitative features of participants in unavoidable crash situations) and rejected (arguing that such an option, if technically feasible, would be discriminatory). The paper's conclusion summarizes its most important insights.

1. Autonomy and justice and why we care about them

Autonomy and justice are two central and mutually related ideas of ethics and political philosophy, as well as the principal values of many people's lives. Although it is difficult, in such a limited space, to even begin portraying their theoretically and practically complex aspects, a selection of their features relevant to the topic of the present paper is presented.[3]

** This work was supported by the Croatian Science Foundation under the project number IP-2022-10-1130.

[1] Cf. e.g. Christoph BARTNECK et al., *An Introduction to Ethics in Robotics and AI*, Cham, Springer, 2021; Markus D. DUBBER, Frank PASQUALE, Sunit DAS (eds.), *The Oxford Handbook of Ethics of AI*, New York, Oxford University Press, 2020.

[2] Such as IEEE GLOBAL INITIATIVE ON ETHICS OF AUTONOMOUS AND INTELLIGENT SYSTEMS, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf (25 March 2024) and UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, Paris, UNESCO, 2022.

[3] The literature on autonomy and justice in ethics and political philosophy is vast. Some valuable introductory readings, also used in conceiving this section of the paper, are John CHRIST-

1.1. Autonomy

Autonomy is an idea with an ancient pedigree and a diversified historical development of its meanings. In Ancient Greece, its usage was almost exclusively political, designating the power of a city-state (*polis*) to pass its own (*auto*) laws (*nomoi*). It was assigned a more individualistic and personal meaning only in the 17th and 18th centuries. For example, the concept of »autonomy of the will« is of crucial importance for Immanuel Kant's deontological ethics focused on duty and unconditional respect for the »moral law«. Without autonomy – interpreted by Kant[4] as the capacity of the will to be motivated exclusively by one's rational considerations – the very possibility of morality would vanish. John Stuart Mill's conception of ethics radically differed from Kant's. As a utilitarian, he maintained that the only criterion of morality of actions is the ratio of their good and bad consequences for the greatest number of people (»greatest happiness for the greatest number«). Nevertheless, autonomy, as a distinctive aspect of individual freedom, was no less important to him. Being the outspoken proponent of modern liberalism, Mill assigned a high value to autonomy, maintaining that a person's happiness would be incomplete if their autonomy were limited contrary to his »harm principle«, which states that »the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others.«[5]

Autonomy occupies a central place in many areas of human life: Parents teach children to think and act autonomously (»deciding for themselves«), even at the expense of making occasional mistakes. When participating in political elections, we like to believe that we autonomously gave our vote to a particular party or candidate despite all the campaigning and advertising around us. When another country invades ours, we see it as an attack on our collective autonomy and freedom that justifies our armed response. When someone claims (in a purely technical context) that a robot, for example, has a certain degree of »autonomy«, many people find this inappropriate and are eager to defend human autonomy as different, unique and indefinitely more valuable.

MAN, Autonomy in Moral and Political Philosophy (29 June 2020), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/autonomy-moral (25 March 2024); Julian LAMONT, Christi FAVOR, Distributive Justice (26 September 2017), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/fall2023/entries/justice-distributive (25 March 2024); David MILLER, Justice (6 August 2021), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/fall2023/entries/justice (25 March 2024); Jonathan WOLFF, *An Introduction to Political Philosophy*, New York, Oxford University Press, 2006.

[4] Cf. e.g. Immanuel KANT, *Groundwork of the Metaphysics of Morals* (translated by M. Gregor), Cambridge, Cambridge University Press, 1997, 41, 57.

[5] John Stuart MILL, *Utilitarianism / On Liberty*, Oxford, Blackwell, 2003, 94.

1.2. Justice

Justice is even more complex and layered than autonomy. »The idea of justice«, according to Miller,

> »occupies centre stage both in ethics, and in legal and political philosophy. We apply it to individual actions, to laws, and to public policies, and we think in each case that if they are unjust this is a strong, maybe even conclusive, reason to reject them.«[6]

Here are some paradigmatic examples of a broad range of justice-related issues that people are concerned about: It is a matter of retributive justice when a person commits a crime (e.g. does not perform the service specified in a contract) and we need to decide about their appropriate punishment. It is a matter of climate and intergenerational justice when we need to figure out which countries and to what extent need to reduce their greenhouse gas emissions to preserve the planet's natural resources for future generations. It is a matter of distributive justice when we need to decide how certain benefits (e.g. public health service) and burdens (e.g. the obligation to pay taxes) are distributed among different members of the society. Since the terrain of distributive justice is relevant for our discussion about the ethics settings of self-driving vehicles, three influential approaches should be mentioned.

According to Robert Nozick's libertarian theory,[7] distributive justice rests upon individual freedom and is adequately served as long as people acquire their property justly and transfer it legitimately to others. Although such a scheme may result in a situation in which some, due to their resourcefulness, own more than others, it would be unjust (an assault on human freedom) if the state intervened and redistributed some of their wealth to those who own less. According to Mill's utilitarian theory,[8] issues of distributive justice – the allocation of benefits and burdens of social life – are best resolved by applying the utilitarian standard that says, roughly, that the right policy is the one that makes the largest number of individuals happy. From the utilitarian perspective, no specific social arrangement is intrinsically the best: depending on the circumstances of a society, any arrangement (from capitalism to socialism) may promote utilitarian justice. According to John Rawls' social-liberal theory,[9] distributive justice is a matter of hypothetical (contractarian) agreement between individuals choosing the principles of justice for a given society without knowing in advance their capabilities and position in it. Rawls argued that anyone deciding behind such a »veil of ignorance« would opt for the following prin-

[6] Miller, *Justice…*
[7] Cf. Robert NOZICK, *Anarchy, State, and Utopia*, Oxford, Blackwell, 1974.
[8] Mill, *Utilitarianism / On Liberty…*
[9] John RAWLS, *A Theory of Justice* (revised edition), Cambridge, Harvard University Press, 1999 ([1]1971).

ciples: (a) that everyone should have the same basic liberties like the freedom of political assembly or religion, (b) that socioeconomic inequalities should be permitted as long as everyone has the same opportunity of gaining wealth or power and (c) that inequalities should be arranged to be to the greatest benefit to the worse-off members of the society.

1.3. The tension

The relationship between autonomy and justice is rich in tension. There is often a trade-off between autonomy and justice, as enhancing one may require deenhancing the other. My autonomous decision to drive 200 km/h through the city center is limited by law to protect the lives of other traffic participants. My autonomous (capitalist) incentive to buy all the gas stations in the country is limited by the law to prevent monopoly and maintain market competition. Physically disabled persons are permitted (they enjoy increased autonomy) to have flexible work hours, compensating thus for other disadvantages they experience in comparison to other workers. These are more or less uncontroversial examples of balancing personal autonomy and interpersonal justice. There are more controversial examples, like free speech vs. the right not to be offended or harmed, affirmative action vs. the fundamental equality of all citizens, drug prohibition or compulsory vaccination vs. the right to control one's body (to mention only some of the most contentious). In the following, a tension between autonomy and justice in the context of self-driving vehicle ethics settings is analyzed.

2. Self-driving vehicles and their ethics settings problem

Imagine that a group of engineers designs a self-driving vehicle of such reliability that it brings about several benefits once introduced into the traffic in sufficient numbers. Compared to conventional cars, self-driving vehicles are not only a step forward when it comes to things like increased mobility for people with disabilities, protection of the environment or car theft prevention – they also, and most importantly, improve traffic safety and reduce the number of traffic fatalities. For the sake of argument, assume that introducing self-driving vehicles reduces the number of fatalities from 100 to 50 annually. Passengers of such vehicles would, by definition, have absolutely no autonomy when it comes to actual driving, but this should be compensated by the overall reduction of traffic fatalities. Imagine further that our group of engineers comes up with a new proposal. They can equip self-driving vehicles with specific ethics settings to prevent even more traffic fatalities (assume that 25 addi-

tional lives are saved). The story is well known: In unavoidable crash situations, ethics settings could distribute the number of fatalities in various ways, e.g. by swerving the vehicle into one pedestrian to save three passengers, by swerving the vehicle with one passenger into the wall to save a group of pedestrians, etc. In such situations, human autonomy (or decision-making capacity) would also be non-existent because the vehicle's behavior would be under its ethics settings control (remember, however, that in unavoidable crash situations, human drivers in conventional cars also usually have 0% autonomy because they lack the time and self-control for any rational reaction). Since self-driving vehicles, as presumed, will have a much better reaction time than human drivers, how should their ethics settings be programmed?

The above question attracted enormous scholarly attention.[10] There can be no doubt that saving the additional 25 lives is as priceless as saving the initial 50 (many would say saving even one life is priceless). However, the problem is that in unavoidable crash situations involving self-driving vehicles with ethics settings, saving specific lives will require sacrificing other specific lives. Whose lives should be saved and whose sacrificed is a thorny issue, with different proposals coming from different theoretical camps. The debate is still intense, but some of the most frequent proposals can be summarized as follows.

2.1. No ethics settings (NES)

Self-driving vehicles should have no ethics settings, i.e. no preprogrammed rules allowing them to choose between human lives. Self-driving vehicles should be as safe as possible for their passengers and all the other traffic participants, but they should not have the option of balancing human lives in unavoidable crash situations. As for the reasons for rejecting the very idea of ethics settings, they can be purely technical, in the sense that no artificial intelligence system will ever be sophisticated enough for such complex decision-making. However, they can also be ethical in that such a decision-making system would violate

[10] Cf. e.g. Edmond AWAD et al., The Moral Machine Experiment, *Nature*, 563 (2018) 7729, 59-64; Bartosz BROZEK, Marek JAKUBIEC, On the Legal Responsibility of Autonomous Machines, *Artificial Intelligence and Law*, 25 (2017) 3, 293-304; Giuseppe CONTISSA, Francesca LAGIOIA, Giovanni SARTOR, The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law, *Artificial Intelligence and Law*, 25 (2017) 3, 365-378; Ryan JENKINS, David ČERNÝ, Tomáš HŘÍBEK (eds.), *Autonomous Vehicle Ethics: The Trolley Problem and Beyond*, New York, Oxford University Press, 2022; Jason MILLAR, 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, 2017, 20-34; Sven NYHOLM, The Ethics of Crashes with Self-Driving Cars: A Roadmap, I, *Philosophy Compass*, 13 (2018) 7, https://doi.org/10.1111/phc3.12507; Sven NYHOLM, The Ethics of Crashes with Self-Driving Cars: A Roadmap, II, *Philosophy Compass*, 13 (2018) 7, https://doi.org/10.1111/phc3.12506.

some fundamental deontological[11] and utilitarian principles.[12] A disadvantage of this option lies in its failure to prevent a certain number of deaths that could have been avoided.

2.2. Personal ethics settings (PES)

Self-driving vehicle passengers should be free to choose their ethics settings. The attraction of this solution is that it preserves one's autonomy to decide how they want their vehicle to react in unavoidable crash situations, which can be in the range from entirely egoistic and always sacrificing pedestrians to entirely altruistic and always sacrificing the passenger(s). The problem with it, however, is the high probability that most people would choose the egoistic ethics settings due to their desire to save their lives or the lives of their near and dear traveling with them. The outcome would be an increased number of traffic fatalities, especially on the pedestrian side. Most people's incentive to choose such extremely selfish ethics settings is not just an assumption: some car manufacturers, for example, have issued statements (which they later withdrew due to negative public response) that their self-driving vehicles will always protect the lives of their passengers at the expense of the lives of pedestrians.[13]

2.3. Mandatory ethics settings (MES)

Mandatory ethics settings – proscribed and enforced by the state – would be such as to save the greatest number of lives. MES would impartially distribute fatalities between those inside the vehicle and those outside the vehicle, giving preference to the side on which more lives could be saved. For example, if forced to choose between saving three passengers and two pedestrians, MES would choose the former; if forced to choose between one passenger and two pedestrians, MES would choose the latter. In terms of its effects, MES can be considered a utilitarian solution (aiming for the best consequences for the greatest number). In terms of its justification, it is also possible to present it as

[11] Cf. BUNDESMINISTERIUM FUR VERKEHR UND DIGITALE INFRASTRUKTUR – ETHIK-KOMMISSION, Automatisiertes und vernetztes Fahren (20.06.2017); https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf (25.032024).

[12] Cf. Tomislav BRACANOVIĆ, No Ethics Settings for Autonomous Vehicles, *Hungarian Philosophical Review* 63 (2019) 4, 47-60.

[13] Cf. Nyholm, *The ethics of…*

a contractarian solution.[14] According to Gogoll and Müller,[15] for example, everyone should realize that it is in their best interest that all self-driving vehicles have MES because it reduces the total number of traffic fatalities and, consequently, one's chances of dying in traffic. However, since not all self-driving vehicle passengers can be trusted to choose the same (impartial) ethics settings, the state must intervene and make such settings mandatory. The presupposed advantage of this solution is the reduction of traffic fatalities and their more just distribution compared to PES. In addition to being unable to prioritize one's life, its possible disadvantage is that it will be an imposed solution that deprives the passengers of their autonomy to control the behavior of their vehicle for themselves.

2.4. An outline of the argument

It will be argued that MES is an unacceptable solution to the ethics settings problem of self-driving vehicles because it violates one elementary principle of justice.[16] The structure of the argument is the following: (a) By relying only on the number of saved or sacrificed lives of passengers and pedestrians, MES violates the principle of justice according to which distribution of burdens to those participating in a given social institution should reflect the distribution of the benefits they derive from that institution. (b) A possible reply to this argument is to make MES more complex, allowing it to rely not only on the number of lives to be saved or sacrificed but also on certain qualitative features of those lives. Since (b), in addition to its probable technical unfeasibility, would come all too close to unjustified discrimination, the conclusion is that (c) MES should be rejected and that self-driving vehicle traffic needs to be such as not to require any similar normative solutions.

[14] Jan GOGOLL, Julian MÜLLER, Autonomous Cars: In Favor of a Mandatory Ethics Setting, *Science and Engineering Ethics*, 23 (2017) 3, 681-700; Derek LEBEN, A Rawlsian Algorithm for Autonomous Vehicles, *Ethics and Information Technology*, 19 (2017) 2, 107-115; Julian MÜLLER, Jan GOGOLL, Should Manual Driving be (Eventually) Outlawed?, *Science and Engineering Ethics*, 26 (2020) 3, 1549-1567.

[15] Gogoll, Müller, *Autonomous Cars…*

[16] The argument developed here complements the one proposed in Bracanović, *No ethics settings…* However, whereas that paper criticizes MES from the ethical perspective, the present criticism presupposes the perspective of political philosophy. A persuasive argument for the thesis that the discussion about ethics settings of self-driving vehicles fits more naturally in the framework of political philosophy instead of ethics can be found in Javier RODRÎGUEZ-ALCAZAR, Lilian BERMEJO-LUQUE, Alberto MOLINA PÉREZ, Do Automated Vehicles Face Moral Dilemmas? A Plea for a Political Approach, *Philosophy & Technology*, 34 (2020) 1, 811-32.

3. »Treating unequals equally« objection to MES

To demonstrate the incongruency of MES with justice, we can rely on one of its elementary principles formulated already by Aristotle in his *Politics* (1131a19-24):

> »So what is just requires at least four terms: the persons for whom it is just are two, and the shares in which its justice consists are two. There will be the same level of equality between persons as between shares, because the shares will be in the same ratio to one another as the persons. For if the persons are not equal, they will not receive equal shares; in fact, quarrels and complaints arise either when equals receive unequal shares in an allocation, or unequals receive equal shares.«[17]

Aristotle's principle, usually abridged as »equals should be treated equally and unequals unequally«, is a commonsensical principle consistent with all three mentioned theories of distributive justice. Here are some of its illustrations: Female teachers are as competent as male teachers and it would be unjust if schools hired only male teachers; since no relevant differences between donut shops and pretzel shops exist, it would be unjust if the state taxed them differently (both instances of the injustice of treating »equals unequally«). Older children are more rational than younger and it would be unjust to discipline them the same for their misbehavior; if student A has excellent grades, whereas student B is mediocre, it would be unjust to write them the same letter of recommendation (both instances of the injustice of treating »unequals equally«). Another suitable example of the functioning of the »equals equally and unequals unequally« principle is special retirement provisions for jobs that are so dangerous or exhausting that no one can perform them for more than a limited number of years (such as police officers, firefighters or paramedics). Those who perform such jobs take upon them a larger share of the burdens needed for the functioning of society, but they are usually compensated for this with adequate retirement benefits (for example, 12 months of their work is converted into 16 months, allowing them to retire earlier). Since they are not equals regarding their burdens, it is only just that they are not equals regarding their benefits. In short, to avoid »quarrels and complaints« about the unjust distribution of benefits and burdens in various social interactions, we must consider relevant and ignore irrelevant differences between individuals participating.

How does MES fare vis-à-vis the »equals equally and unequals unequally« principle? It might seem that it fares well because it treats self-driving vehicle passengers and pedestrians equally, impartially distributing among them the risk of being killed. If more lives can be saved by sacrificing pedestrians, it will sacrifice the pedestrians, and if more lives can be saved by sacrificing

[17] Aristotle, *Nicomachean Ethics* (translated by R. Crisp), Cambridge, Cambridge University Press, 2004, 86.

the passengers, it will sacrifice the passengers. When the unavoidable crash situation occurs, the MES's distribution of its burdens (the likelihood of being sacrificed) and benefits (the likelihood of being saved) is equal, as the only unit of its decision-making is the number of lives. However, MES violates the »equals equally and unequals unequally« principle from another angle. Considered as groups, the self-driving vehicle passengers and pedestrians seem in several respects sufficiently unequal so as not to be treated equally in unavoidable crash situations. Considering the broader context of self-driving vehicle traffic, MES's equal distribution of burdens among its participants (in terms of the same likelihood of being killed in unavoidable crash situations) has no adequate justification in the prior (unequal) distribution of benefits they derive from that traffic. Three such inequalities are outlined below.

3.1. Unequal advantages

Self-driving vehicle passengers and pedestrians will most likely be unequal regarding their advantages from self-driving vehicles. Just like owning or driving a conventional car, all other things being equal, is an advantage to walking, so is owning or being a passenger in a self-driving vehicle an advantage to being a pedestrian. Such a vehicle enables one not only to get quicker from point A to point B and thus save time (and money) but also to pursue many life-important activities, like having a business meeting or a romantic date, watching a movie, reading a book, preparing for philosophy exam or simply taking a nap. Most of these activities remain unavailable even to conventional drivers, let alone pedestrians. That a person can afford a self-driving vehicle also strongly suggests that she has a higher socioeconomic status than someone who needs to walk or ride a bike.[18] Theories of distributive justice, like social-liberal theory or some versions of utilitarian theory, might consider this difference substantial enough to justify an unequal distribution of burdens among self-driving vehicle passengers and pedestrians (favoring the latter) in unavoidable crash situations. By focusing on reducing crash fatalities based only on the number of individuals involved, MES remains insensitive to this (from the justice perspective) relevant difference.

[18] For a similar suggestion cf. Keith ABNEY, The Ethics of Abuse and Unintended Consequences for Autonomous Vehicles, in: Jenkins, Černý, Hříbek (eds.), *Autonomous Vehicle Ethics*…, 258-274; Robert SPARROW, Mark HOWARD, Make Way for the Wealthy? Autonomous Vehicles, Markets in Mobility, and Social Justice, *Mobilities*, 15 (2020) 4, 514-526.

3.2. Unequal exposure to danger

Pedestrians are »particularly vulnerable« because »they do not wear protective clothing and, compared to other transport modes, they have a low and vulnerable position on the road« and, as a consequence, »almost all fatalities (98%) in pedestrian crashes are the pedestrians themselves.«[19] Although self-driving vehicles with MES promise to reduce the overall number of traffic fatalities, it is reasonable to assume that pedestrians will continue being the most vulnerable party in traffic. For example, self-driving vehicles will probably be mutually connected and coordinated (and thus additionally protected) in unavoidable crash situations, but no similar connectedness and coordination will take place between self-driving vehicles and pedestrians. MES algorithms will probably not be able to address all problematic situations optimally (e.g. when it will be difficult to assess the exact level of risk for all participants of a potential crash), and in such scenarios, being a self-driving vehicle passenger will be preferable to being a pedestrian. There is also the difference between passengers and pedestrians regarding their statistical exposure to danger. Assuming a self-driving vehicle passenger and a pedestrian need to travel the same distance (e.g. from A part of the city to B part of the city), the pedestrian will spend more time in traffic than the passenger and interact with more traffic participants (conventional cars, self-driving vehicles and pedestrians). The probability of him being involved in a possible accident will be higher than the probability of a self-driving vehicle passenger. In other words, from the perspective of exposure to danger, self-driving vehicle passengers and pedestrians are unequal and MES should not treat them equally.

3.3. Unequal responsibility

A peculiar outcome of introducing self-driving vehicles with MES into the traffic would be creating two different groups of traffic participants regarding their autonomy and, consequently, their legal and moral responsibility. The pedestrians would retain autonomy to participate in traffic more or less the same way as before, including their current legal and moral responsibility for their actions and omissions (e.g. jaywalking or failure to yield to emergency vehicles). Self-driving vehicle passengers, practically by definition of MES, would be deprived of their autonomy and, consequently, their legal and moral responsibility for anything in traffic involving their vehicles. Their role would be reduced to instructing the vehicle where they wish to go, while its manufacturer (or perhaps the state that proscribed MES) would take full responsibility for

[19] EUROPEAN ROAD SAFETY OBSERVATORY – DIRECTORATE GENERAL FOR TRANSPORT, *Facts and Figures – Pedestrians*, Brussels, European Commission, 2021, 5.

possible damages resulting from its (mal)functioning or (mal)programming. In this context, not having autonomy is desirable because one can enjoy the benefits of autonomous driving without worrying about being responsible for anything that might go wrong. This is another instance of the unequal (and probably unjust) distribution of benefits and burdens among self-driving vehicle passengers and pedestrians: the likelihood of being killed in a potential crash situation is distributed among them symmetrically, but their legal and moral responsibility preceding such situations in traffic is not.

3.4. Reasons for rejecting MES

The argument against MES can be summarized as follows: By restricting human autonomy (in terms of both driving the vehicle and decision-making in unavoidable crash situations), MES promises to reduce the number of traffic fatalities. This transfer of autonomy from human drivers to self-driving vehicles and their ethics settings comes with a cost in terms of justice. All participants in unavoidable crash situations are treated by MES equally, despite minimally three significant inequalities with which they enter such situations: the inequality of advantages from self-driving vehicles (including the inequality of access to them), the inequality of (physical and statistical) exposure to danger in traffic, and the inequality of legal and moral responsibility.

Imagine four *bon vivants* who go to crazy parties every night, returning home in the early morning hours in a self-driving vehicle with MES. If their vehicle finds itself in an unavoidable crash situation with a group of three construction workers going to work, *bon vivants*' lives would be prioritized because they outnumber the workers. This does not resonate well with our intuitions about justice because the participants of this situation found themselves in the same place at the same time for different reasons (going to work is a necessity, attending parties a luxury), the danger they are exposed to is different (being a passenger in a self-driving vehicle is safer than being a pedestrian) and their legal and moral responsibility is different (*bon vivants* bear no responsibility for their self-driving vehicle actions whereas construction workers remain responsible for their behavior). MES, therefore, as long as it focuses only on the number of lives saved in unavoidable crash situations, treats »unequals equally« and is an unjust solution for reducing traffic fatalities.

4. A possible reply: MES 2.0

Under the assumption that the above argument is sound, a way to save MES could be to make it more complex (we can call it MES 2.0). MES 2.0 could be programmed, let us imagine, to rely not merely on the number of individual lives to be saved or sacrificed in unavoidable crash situations but also on some of their additional qualities, such as age, gender, health status, number of dependents, etc. Returning to our example, an MES 2.0 self-driving vehicle could now save three construction workers instead of four *bon vivants* because the number of people who depend on their salaries, let us say, exceeds the number of people who depend on the salaries of *bon vivants*. Or, to use another example, an MES 2.0 self-driving vehicle could save the life of one healthy 20-year-old pedestrian instead of the lives of a seriously ill 90-year-old couple traveling to the hospital for a regular checkup. MES 2.0, in other words, would allow self-driving vehicles to make some life-or-death choices that would probably resonate better with most people's intuitions about justice than choices made exclusively based on the number of potentially saved or sacrificed lives. And yet, MES 2.0 would most likely face some objections that would render it implausible. We can briefly present two of them: technical unfeasibility and discrimination.

4.1. Technical unfeasibility

In discussions about the ethical aspects of various applications of artificial intelligence, it is often assumed that their highly advanced abilities are possible or will become possible in the foreseeable future. There is a good methodological reason for such an assumption: ethical reflection on new technologies often needs to anticipate their potentially harmful development to prevent it from happening while there is still time. However, regarding the debate about ethics settings of self-driving vehicles, the question is how intelligible such an approach is. Are ethics settings – such as those appearing in various thought experiments – technically feasible? How likely is it that someday, self-driving vehicles will come equipped with supercomputers that will crunch vast amounts of data about a vast number of individuals to instantly, almost divinely, make decisions about the comparative value of their lives in potential crash situations? Probably even more fundamental is the following question: If self-driving vehicles of the future are going to be technologically advanced as is assumed by MES or MES 2.0, isn't it then also rational to assume that the entire traffic will be organized in such a technologically advanced and safe way that the issue of ethics settings of self-driving vehicles will actually become a non-issue?[20]

[20] Other problems are also possible. For example, a question could arise – discussed in Bracanović, No ethics settings… – about the storage and access to personal data necessary

4.2. Discrimination

Imagine that MES 2.0 becomes technically possible and can distinguish participants in unavoidable crash situations according to age, gender, health status, number of dependents, etc. Would that solve the problem of focusing only on the number of lives and ignoring their other (from the perspective of justice) significant differences? This could solve that problem to some extent but at the cost of exposing MES 2.0 to an equally severe objection of unjustified discrimination. The point is that if one group of people (such as persons of a particular age or health status) were to become systematically profiled as preferred victims in unavoidable crash situations, MES 2.0 would undoubtedly be seen as an instrument of institutionalized discrimination (remember, it is imposed by the state). It would treat separate persons not according to their individual and unique characteristics, but according to the characteristics of their demographic group. Such treatment is typically considered discriminatory as it violates the principle of »moral individualism«, according to which »how an individual should be treated depends on his or her own particular characteristics, rather than on whether he or she is a member of some preferred group«.[21] Of course, it is open for debate whether all instances of such treatment would always amount to unjustified discrimination.[22] However, should the public only start to suspect (which is a likely scenario) that MES 2.0 involves certain discriminatory elements akin to ageism, sexism or racism, that could negatively influence its willingness to accept self-driving vehicles and their potential to improve traffic safety significantly.

Concluding remarks

Transferring our driving autonomy to self-driving vehicles promises to save lives, whereas transferring our autonomy to decide who will be saved or sacrificed in unavoidable crash situations to MES promises to save even more lives.

for the functioning of MES 2.0, including the issue of bias and transparency of its processing. A just as interesting problem – mentioned in Jeffrey K. GURNEY, Unintended externalities of highly automated vehicles, in: Jenkins, Černý, Hříbek (eds.) *Autonomous Vehicle Ethics…*, 147-158 – is that the reduction of the traffic fatalities that will occur thanks to self-driving vehicles might entail a reduction of the organs available for transplantation. When deciding whom to sacrifice in an unavoidable crash situation, should MES 2.0 also consider the number of people waiting for organ transplantation?

[21] James RACHELS, *Created from Animals: The Moral Implications of Darwinism*, Oxford, Oxford University Press, 1990, 5.

[22] Cf. Vincent CHIAO, Algorithmic Decision-Making, Statistical Evidence and the Rule of Law, *Episteme*, online first, https://doi.org/10.1017/epi.2023.27; Derek LEBEN, Discrimination in Algorithmic Trolley Problems, in: Jenkins, Černý, Hříbek (eds.), *Autonomous Vehicle Ethics…*, 130-142.

A problem with the latter autonomy transfer, as argued in this paper, is that it distributes burdens (the number of fatalities) among self-driving vehicle passengers and pedestrians equally, disregarding the unequal prior distribution of benefits these two groups have from self-driving vehicle traffic. In other words, MES fails to respect the elementary principle of justice, according to which equals should be treated equally and unequals unequally. Assuming that ethics and political philosophy, as it seems to follow from the analysis of the MES 2.0 problem with unjustified discrimination, are unlikely to find an acceptable normative solution to the problem of ethics settings, it is perhaps better to start looking for technological approaches to reducing fatalities in traffic with self-driving vehicles that will not require any similar life-or-death decision-making procedures.

Tomislav Bracanović*

Autonomna vozila, autonomija i pravednost

Sažetak

Samovozeća vozila, uvriježeno je vjerovanje, morat će biti opremljena takozvanim etičkim postavkama, čime će se smanjiti broj smrtnih slučajeva u situacijama neizbježnih sudara. Ove etičke postavke, također je uvriježeno vjerovanje, trebale bi biti obvezne i nepristrano distribuirati smrtne slučajeve između putnika u vozilu i pješaka, dajući prednost strani na kojoj se može spasiti više života. Ovaj rad kritizira oba ova vjerovanja, usredotočujući se na napetost koju će obvezne etičke postavke stvoriti između vrijednosti autonomije i pravednosti. Središnji argument je da obvezne etičke postavke, sve dok se oslanjaju samo na broj života koji će biti spašeni ili žrtvovani, ne poštuju elementarno načelo pravednosti (»jednake treba tretirati jednako, a nejednake nejednako«), zanemarujući neke značajne razlike između putnika u samovozećim vozilima i pješaka. Nakon odbacivanja mogućeg odgovora na ovaj argument (da bi se obvezne etičke postavke također mogle oslanjati na neke kvalitativne značajke sudionika u situacijama neizbježnih sudara), zaključuje se da nikakvo općeprihvatljivo normativno rješenje za problem etičkih postavki nije izgledno i da prednost treba dati tehnološkim pristupima smanjenju smrtnih slučajeva u prometu sa samovozećim vozilima.

Ključne riječi: autonomija, diskriminacija, jednakost, pravda, samovozeća vozila.

* Dr. sc. Tomislav Bracanović, znanstveni savjetnik, Institut za filozofiju, Ulica grada Vukovara 54, HR-10000 Zagreb; e-mail: tbracanovic@ifzg.hr.