AI and Ethics (2022) 2:177–184
https://doi.org/10.1007/s43681-021-00088-7

ORIGINAL RESEARCH

God does not play dice but self-driving cars should

Luvuyo Gantsho[1]

Received: 13 May 2021 / Accepted: 29 August 2021 / Published online: 8 September 2021

Abstract
Advances and improvements in computing power and processing have led to a clear upward progression in the degree to which autonomous vehicles can operate freely without human involvement. Advances in autonomous vehicle technology may reduce the incidence of vehicle accidents born from human error and would be a general benefit if widely used and properly regulated. However, with increases in machine agency comes the corresponding challenge of machine ethics that must keep pace with the increasing number of decisions autonomous cars need to make. In this paper, I explore and advance a view on how autonomous vehicles ought to respond in a particular tragic choice scenario under a specific set of constraints where any one person needs to die for the sake of many more. I argue that in such cases autonomous vehicles ought to randomly select who to sacrifice and that such random selection ought to be blind to particulars that set people apart from each other, including whether a potential sacrifice is a passenger or owner of the self-driving car in question.

Keywords Machine ethics · AI · Self-driving cars · Tragic choice scenario · Random selection

1 Introduction

Self-driving cars, also known as autonomous vehicles, driverless cars, or robo-cars [1, 2], are vehicles capable of environmental awareness and navigation with little or no human input [3, 4]. Advances and improvements in computing power and processing have led to a clear upward progression in the degree to which autonomous vehicles can operate freely without human involvement. Combination, fusion and integration of various sensory technologies such as radar, lidar, sonar, GPS, odometry and inertial measurement units allow self-driving cars to adequately perceive their surroundings [2, 3]. Advances in control systems allow for more swift and ergonomically efficient interpretation of sensory inputs that have led the way for more seamless adoption of appropriate navigation paths, obstacle aversion and response to traffic signage [3, 5, 6]. Given these advances, it seems highly plausible that self-driving cars will reach a stage where they can operate fully automatically under all roadway and environmental conditions that can be managed by human drivers [7].

The benefits and applications of fully autonomous vehicles cannot be understated. According to a 2020 *Annual Review of Public Health,* if autonomous cars were properly regulated, they would likely reduce morbidity and mortality attributed to vehicle accidents and may help reshape city planning to optimise healthy urban environments [8]. The same review concludes that if 90% of cars in the United States of America became fully autonomous, an estimated 25 000 lives would be saved annually which equates to over $200 billion annually. According to another 2020 study, fully autonomous cars would increase productivity and housing affordability, as well as reclaim land that is currently used for parking [9]. Given the above-mentioned statistics, it is safe to assume that self-driving cars would be an overall benefit to society if their use is permitted and properly regulated. While I conform to the view that, generally speaking, technological advances that produce more benefits than harms are permissible [10], I submit one way in which the widespread use of fully autonomous cars is morally problematic. The decisions that fully autonomous cars need to make in impending fatal traffic collisions.

The girth of the ethics debate on decisions fully autonomous cars need to make in catastrophic situations is vast. From who should be held liable in car accidents involving fully autonomous vehicles [11]; how self-driving cars ought to navigate trolley problem scenarios that pits utilitarian

✉ Luvuyo Gantsho
lgantsho@yahoo.com

[1] University of Witwatersrand, Johannesburg, South Africa

commitments of saving the most people in catastrophic scenarios with deontological commitments to not harm or sacrifice persons [12]; the ethical distinction between an autonomous car killing an innocent bystander and letting an innocent bystander die [13] or whether autonomous vehicles should prioritise the protection of passengers over pedestrians [14]. While the above-mentioned ethical quandaries are valid concerns that if left unresolved will continue to obfuscate our ability to programme coherent moral algorithms into autonomous vehicles, I hope to bring to the fore in this paper one moral resolution that would aid in bringing us closer to a coherent moral framework for autonomous vehicles.

Attempting to resolve all the possible moral dilemmas that might arise from the emergence of autonomous vehicles is a herculean task that I admit from the onset I make no attempt at addressing *in toto*. I humbly only seek to argue for one specific outcome whose resolution I hope is sufficiently compelling enough to not allow anything in way of controversy with respect to how autonomous vehicles ought to respond in that specific scenario. The research question I consider in this paper is whether autonomous vehicles ought to be programmed to randomly sacrifice *one* person in tragic choice scenarios where any one person can be sacrificed for the sake of *many* more persons. I argue that autonomous vehicles must randomly select who must be sacrificed in situations where any one person needs to be sacrificed for the sake of many more persons. I argue that the process for random selection must be blind to *all* characteristics that set persons apart from one another, including who owns or is a passenger inside the autonomous vehicle involved.

There are a number of limitations and presumptions I argue from in this paper that on one hand, I admit, trivialises my discourse by limiting my findings to a small set of catastrophic scenarios. However, on the other hand by so doing I hope that the elimination of several moral confounds makes my thesis sufficiently plausible. The first limitation that I stipulate in my thesis statement is that the number of people that need to be sacrificed needs to be limited to a single person, while the number of people that stand to be saved is greater than one. This limitation effectively rules out disruptive utilitarian counter arguments that would make uncontroversial claims that autonomous vehicles should randomly select who to sacrifice more difficult especially in cases where a large number of people stand to be sacrificed for a smaller number of people. A second limitation is that I will only consider catastrophic scenarios where all participants concerned are innocent of any wrongdoing. This second limitation eliminates desert-based claims that might make arriving at a coherent conclusion challenging if desert-based claims favour certain people for sacrifice over others, thereby making an argument for random selection more controversial.[1] A third limitation is that I will only consider catastrophic scenarios where *human persons* (as opposed to non-person humans or non-human non-persons viz. animals) are involved. This third limitation eliminates contrary theories of moral status that either do not advance a plurality of different moral statuses or advance an expanded definition of personhood that may include nonhumans. By limiting my discourse to human persons, I do not have to contend with controversies pertaining to whether severely mentally deficient humans or animals should be given the same consideration as "normal" humans.

In what follows in this paper, my argument will be divided up into three parts. In the first part, I argue for and defend a plausible theory of moral status that confers equal status to all beings that fit the definition of persons. In the second part, I argue that the equal moral status of persons confirms a similar *equality from inviolability of moral rights.* That all person's interests must be given equal consideration (these interests include equality from harm or death). I argue in line with one author that this equality from inviolability is best expressed through some randomised nondeterministic programming that selects who to sacrifice in cases where persons are involved in catastrophic scenarios where any one person need only be sacrificed for the sake of many more. In the third part, I defend this programming in this scenario in cases where autonomous vehicles might sacrifice the owner of or passengers in autonomous cars and cases where age, health or intelligence or wealth of participants in catastrophic scenarios are characteristics that set potential sacrifices apart. I conclude that *in addition to whatever else the programming of autonomous cars ought to be*, autonomous cars ought to be programmed to be indifferent in their random sacrificial selection of human persons involved in catastrophic scenarios where any one person needs to be sacrificed for many more human persons. This view ought to hold even if it leads to the owner or passenger of an autonomous car gets sacrificed or if the person being sacrificed is much younger, healthier, wealthier, or smarter than the persons be saved.

[1] By desert-based claims, what I am referring to is cases where select participants in catastrophic scenarios might be deserving of retribution, punishment or justice to the effect that certain individuals should possibly be preferred as sacrifices over others. For example, a person has created a catastrophic scenario by breaking the rules of the road through jay walking or not wearing one's seat belt or alternatively has found themselves in a catastrophic scenario after robbing a bank.

2 Plausible theory that confers moral status

In this section, I advance and defend a theory that accords moral status. I argue that a Kantian theory that treats moral status as a threshold concept provides the most plausible theory that accords moral status. This is in contrast to utilitarian theories that confer variable moral considerability, that while coherent, advance prescripts that lead to far-reaching absurd outcomes that are also generally misaligned with most people's intuitions. I conclude that given the sensible prescripts and intuitive appeal of the Kantian approach, we should endorse the view that moral status is best thought of as a threshold concept. That everyone who has the characteristics and capabilities that confer personhood are persons of equal moral status.

An entity or being is said to have 'moral status' if and only if its interests morally matter to some extent for the entities or being's own sake. For example, an animal may have moral status, if its suffering ought to be taken into consideration on account of the animal itself, and regardless of other beings [15]. The term moral status and moral standing are typically used interchangeably. However, for this paper, I will follow after Allen Buchanan's [16] definition that a being has *moral standing* if it counts morally, in its own right. In contrast to the notion of moral standing, *moral status* is a comparative notion, that is, two beings can have moral standing, but one might have a higher moral status than the other. Moral theories that accommodate a plurality of different moral statuses often regard humans—or at least human *persons*—as beings with the highest moral status. In this regard, persons can be thought of as possessing the highest extant moral status. And while I do not necessarily advance this view, the reason why most people would advance the view that humans (or at least human persons) should be favoured over animals (or at least nonperson animals) in catastrophic one-on-one scenarios is perhaps because humans are persons while animals are not.

In an attempt to distinguish between the moral statuses of humans and animals, Buchanan [16], after McMahan [17], recounts two philosophical theories that seek to explain the difference in moral status between persons and animals: interest-based accounts, and respect-based accounts.

2.1 Interest-based accounts for moral status

According to the interest-based account, the moral status of a being ought to depend, roughly, on how much good its life involves. 'Good' in this context refers to the well-being of the being in question, which is understood as comprising of various interests. This view implies, amongst other things, that how wrong it is to kill that being depends on how much well-being that being stands to lose by dying.

According to Buchanan [16], it is not clear that the interest-based account of moral status can give a robust enough support for the folk view that human persons have a *much* higher moral status than non-human non-persons (animals). Buchanan recommends that the interest-based account, if properly understood, actually discredits the notion that there are different moral statuses, instead it recommends replacing the concept of different moral status with '*a continuum or gradient of moral considerability*'. One can image various beings with different interests that range seamlessly from those that are most bountiful (the highest conceivable goods, such as a human person self-actualizing or becoming a grandparent), to those that are incredibly marginal (the lowest goods such as a fly cleaning its wings). Because the interests a being can have both between and within species are spread so seamlessly, the interests-based account does not seem to support the thesis that there is a discreet threshold of moral status that equally encompasses and substantially elevates all human persons above animals (or at least those animals that uncontroversial have less moral status than humans such as rats, cockroaches and single celled organisms). The concept of moral status, as our folk intuitions would suggest, consists of discreet thresholds that are incompatible with the interest-based accounts continuum of moral considerability.

This is not to say that the interest-based account with its continuum of moral considerability cannot intuitively motivate for the moral status between humans and animals in most cases (in fact it can[2]). Instead, I argue that the interest-based account with its continuum of moral considerability cannot account for why humans *always* have a higher moral status than animals or for why the interests-based account supports the view that some humans have a higher moral status than other humans. This is because while it can be easily argued that humans *generally* have higher interests than animals, this will not always be the case, and it is certainly the case that some people have higher interests than other people. The interest-based account would seem to suggest (rather counter intuitively) that in some select cases, an animal can have a higher moral status than a human person if its interests are high enough and the human persons interests are low enough and that humans with lower interests have lower moral status. So while an appeal to the interest-based account can intuitively support the view that humans have a higher moral status than animals in most cases, it cannot motivate for the broader theoretical (and intuitive) position that human persons *always* have a substantially higher moral status than animals or that humans are moral equals. Motivating from the interest-based account would

[2] Viz. humans have a higher moral status than pigs, because humans relish in the arts, while pigs enjoy nothing more than a mud bathing.

back its proponents into an uncomfortable corner wherein they would have to concede that it is at least *possible* for an animal to have a higher moral status than a person and that some humans are morally superior to other humans.

2.2 Respect-based accounts for moral status

According to the respect-based account of moral status, which is grounded in Kantian moral philosophy, all beings that possess certain capacities are conferred personhood. On this view, there is no room for the concept of a continuum of moral considerability that the interest-based accounts espouses [16]. Many contemporary social contractarian moral theorists hold that beings who have the capability and interest to engage in mutual accountability through the giving and heeding to reason ought to be considered persons *regardless* of how good they might be at engaging in acts of mutual accountability [18, 19]. Buchanan notes that the primary challenge that the respect-based account of moral status faces is the ambiguity of where it is supposed threshold should be placed. It may be difficult to judge whether some human beings (e.g., babies or cognitively impaired adults) have the necessary capabilities for mutual accountability. Nevertheless, Buchanan notes, the respect-based account is still effective insofar as we can identify cases of persons who unambiguously meet the requirement of being capable for mutual accountability. More importantly, he adds that the respect-based account is able to articulate why anyone who meets the threshold in question has an equal moral status to anyone else who meets it. There is no room for degrees of moral considerability on the contractualist or on the Kantian understanding of the respect-based view because both understandings confer personhood to anyone that has the capability and interest for practical rationality and/or mutual accountability. Buchanan adds that at least on the social contractualist-based view for the conferment of moral status, the grounds for which moral personhood is conferred—the capability and interest for mutual accountability—also serve as the source for moral principles that motivate for the recognition of moral status appropriately accorded to persons.

When compared to one another, the respect[3]- and interest-based accounts are distinguished mainly by the former endorsing a threshold view of moral status, while the latter endorses a continuum-based view for moral considerability. Of these two views, it is the respect-based account with its threshold conception of moral status that aligns most closely to folk intuitions concerning how moral status is accorded. This is because the the respect-based view (in contrast to the interest-based view) asserts that all beings that qualify as persons possess moral statuses that are equally and substantially elevated over none persons.

Buchanan [16] adds that there is a more fundamental difference between the respect- and interest-based accounts. The former is primarily concerned with *persons* while the latter is primary concerned with *interests.* Buchanan argues, that the respect-based view is the more plausible account for how moral status ought to be accorded, not only because it endorses the readily accepted threshold concept, but also because it is precisely concerned about persons (or the capacities and interests that make them thus). The interest-based account focuses on interests as the principal object of moral concern, not persons. Buchanan argues that in this regard, it is committed to the odd view that persons do not have moral status in and of themselves; that the phrase 'persons have moral statuses higher than animals' is to the adherent of the interest-based view better phrased as 'certain interests are so morally important that it is appropriate to treat those beings whose interests they are as if they have a higher moral status'. As such, the interest-based view is perplexing for two reasons: (1) it endorses the counter intuitive view that moral status is better described as a continuum of moral considerability[4] and not as a threshold, and (2) the interest-based account holds that it is the interests that are held by a being that as so morally important as to ground how moral considerability ought to be accorded, not the being's capability and interest to act on reason and mutual accountability. It is for these two reasons that Buchanan concludes that if it is requisite for a plausible theory of moral status to be intuitively appealing then the interest-based account ought to be outweighed by the respect-based account.

I argue farther and assert that an adoption of the interest-based account with its continuum of moral considerability leads to far reaching societal and legal implications. That because human rights are an articulation of moral status and, therefore, the respect-based account [20], then an endorsement of the interest-based view would be to reject the concept of human rights in favour of a gradation of legal privileges that are awarded on the richness of a being's interests. In such a world, people would gain and lose legal privileges on the basis of whatever fleeting interests they decided to adopt. Given how far reaching an impact the concept of human rights has had on human civilization and how massively positive a role it has played in improving the lives of billons, to abandon that in favour of the interest-based account with its gradation of legal privileges can only be described as *absurd.* As such, given the interest-based

[3] For the purposes of brevity, any future reference I make to the respect-based account also includes the contractualist based view.

[4] That leaves the door open to the possibility of some animals having a higher moral status than some humans, or some humans having a higher moral status than other humans.

account's absurdity, it should be rejected in favour of the respect-based account with its endorsement of human rights.

3 Inviolability and the moral status of persons

Having settled on the respect-based account of moral status with its readily acceptable threshold concept, I now consider an important implication that follows from viewing moral status as a threshold concept. That because persons ought to be thought of possessing identical moral statuses and because the conferment of moral status implies inviolability of said status, then whatever rights, benefits and privileges that personhood confers ought to be equally inviolable for all persons. I then argue in line with one author that person's equality of inviolability is best expressed through programming autonomous vehicles to randomly and non-deterministically decide who to sacrifice in cases of catastrophic scenarios.

Following from the previous section, because moral status ought to be thought of as a threshold concept then it ought to simply follow that the personhood that moral status confers is *inviolable.* That is, the rights, privileges and benefits that confer personhood cannot be violated for someone else's sake. To violate the rights, privileges and benefits that personhood confers for the sake of other persons would erode at the very core of that person's moral status. It would be absurd to consider two beings as being persons of equal moral status and yet one person can *rightfully* be sacrificed for the sake of another person. To forsake one person for another would be a flagrant disregard of their equal moral status. Therefore, the inviolability of moral status, when properly understood, is also a threshold concept, such that all the benefits and privileges that come with personhood are inviolable irrespective of how well a person can reason or engage in mutual accountability—whatever rights and benefits that a person has cannot be taken from them for the sake of other persons.

However, in line with Buchanan [16], even if one concedes the validity of inviolability such a view has practical limits. In tragic choice situations the rights of a person (including their right to life) can be permissibly infringed to avert the death of many in cases where any one need only die. Buchanan motivates for an exception to inviolability in 'supreme emergencies' a concept similar to my catastrophic scenarios.

> Perhaps a plausible understanding of the inviolability of persons should allow for the possibility that even the most basic rights of persons, including even the right to life, can be infringed in a supreme emergency, a tragic choice situation where the deaths of a great many innocent persons can be avoided only by killing a few innocent persons

Given the equality of inviolability shared across all persons, any exception to the rule of inviolability such as sacrificing one to spare the many which would permissibility be the case in catastrophic scenarios, ought to be a sacrificial exemption that equally applies to all persons. This understanding of inviolability affirms that even if a *person* may be sacrificed for the lives of many *persons*, the choice of deciding to sacrifice one for the sake of the many should not be made on the false premise that some persons are inferior to other persons (as many people's prejudices may incorrectly have them believe). On the view of equal moral status, it seems more plausible to choose who ought to be sacrificed by some fair lottery system. The threshold view of inviolability excludes taking the well-being a person stands to lose by dying, into consideration [16]. Therefore, in the context of self-driving cars that find themselves in similar tragic choice scenarios where any one person can be sacrificed to save many more, the morally appropriate computer programming that self-driving cars ought to have must involve some random non-deterministic system that selects who to sacrifice. This random non-deterministic programming would adequately replicate a fair lottery system that on the view of equal moral is the appropriate way to select who to sacrifice in tragic choice.

In my argument for the equal inviolability of persons, I have conceded that the inviolability of a person's rights can be violated in dire situations where tragic choices of life and death must be made. I argued that the best way to reconcile the view that inviolability of moral status is a threshold concept with the view that persons can be sacrificed for others in catastrophic situations is for autonomous vehicles to select who ought to be sacrificed with a random nondeterministic programme. Taking any capability, interest or good into consideration when deciding which person should die would be to contradict the respect-based account with its threshold concept of moral status and inviolability. Some innocent persons may be sacrificed for the many, but such a decision must be blind to any of the differences that set persons apart from each other.

4 Implications for a blind lottery-based system on who might end up getting sacrificed

Having motivated for why autonomous vehicles should randomly and non-deterministically select which persons ought to be sacrificed in catastrophic scenarios where any one person needs to be sacrificed to save the lives of many more, I am now poised to examine some notable implications this

has on how autonomous cars ought to behave in catastrophic scenarios. The implications I consider is what role personal particulars, such as age, wealth, intelligence, health and who is passenger or owner of the autonomous vehicle doing the sacrificing should factor into autonomous vehicle's decision into who to sacrifice. I argue that drawing *any* distinction between persons on the previously mentioned basis into who should be sacrificed either is an implicit adoption of the interest-based account of moral status or an implicit rejection of the equal inviolability between persons that ought to be upheld.

As previously discussed, the interest-based account with its continuum of moral considerability is a theory that accords the highest degree of moral considerability to beings who stand to lose the most well-being if they died. Intuitively and in line with descriptive analysis of *The Moral Machine Experiment* [21], Global moral preferences have a bias towards who should be sparred in tragic life and death scenarios. According to that experiment there was a global moral preference for sparing young lives over old lives, sparring the healthy over the unhealthy, and sparring the wealthy over the poor. Admittedly, it is not readily apparent what theoretically motivates for these preferences. However, moral legitimacy for these views can be expressed in terms of the interest-based account. Relatively younger, healthier and wealthier people would tend to loss more well-being if they died as compared to older, sicklier, and poorer people. Therefore, on the interest-based account with its continuum of moral considerability, the interests of the healthier, older and wealthier should take precedent over their less fortunate peers. In catastrophic scenarios, the interests of the better off would probably be better served by autonomous cars ranking potential sacrifices in order of who would loss the least amount of well-being by dying and sacrificing those individuals in that order, as opposed to the random nondeterministic system preferred by the respect-based account [16].

However, as already established I gave three reasons why we should not endorse the interest-based account; its non-intuitive advancement for a moral continuum, its peculiar preference for interests over capabilities and the absurd implications it would lead to. We ought to rather endorse the respect-based account with its threshold concept of moral status and inviolability that all ought to have an equal chance of being sacrificed in catastrophic scenarios and this equality is best expressed through autonomous vehicles being programmed to decide in a random and nondeterministic manner who should be sacrificed.

Following on, the respect-based account asserts that moral status ought to be thought of as a threshold concept, this also implies that inviolability ought to be thought of as a threshold concept. In catastrophic scenarios where any one person needs to be sacrificed for the sake of many more people, autonomous cars ought to be programmed to select who is to be sacrificed in a random nondeterministic manner because this manner of selection respects the equality of inviolability that defines personhood. To programme automated vehicles to prioritise the life of the passenger or owner of the self-driving vehicle in catastrophic scenarios would be to reject the equal inviolability that defines personhood. Because we ought to endorse the respect-based view we ought to treat the inviolability of persons as a threshold concept which means in catastrophic scenarios all should have an equal chance of being sacrifice. Programming self-driving cars to prioritise the owner or passenger of automated vehicles would be to assign unequal inviolability between persons which goes against the respect-based account of moral status that we ought to endorse. Because of the respect-based account of moral status, passengers and owners ought not be prioritised in this particular catastrophic scenario. Owners and passengers of autonomous vehicles ought to stand an equal chance of being randomly selected for sacrifice.

While my findings on what the moral programming of autonomous vehicles ought to be are very narrow with respect to the type of tragic choice scenario they seek to resolve, my findings do prove applicable for at least one use case. Imagine a scenario where an autonomous car, containing one passenger, is driving down an overpass at considerable speed that is banking sharply to the right. The car has experienced a catastrophic break failure and cannot stop. There are two pedestrians on the road that are oriented in such a manner that if the self-driving car does not steer from its current trajectory, both pedestrians will die, and the car will proceed to crash into the side railing of the overpass at close to a perpendicular angle, thereby killing the passenger. That is, if the self-driving car does nothing, everyone dies. However, there are three available courses of action available to the self-driving car that, if executed, could save any two people in exchange for any one sacrifice. The car can either steer sharply to the left, saving the two pedestrians and killing the driver due to crashing into the side railing at close to a perpendicular angle. Bank slightly to the left killing the leftmost pedestrian while saving the passenger (due to a shallower approach with the side railing) and the rightmost pedestrian. Or the self-driving car can bank slightly to the right, killing the rightmost pedestrian, while saving the leftmost and the passenger (again, due to the cars shallower approach with the side railing). Few would challenge that the above scenario is at least *possible*, and if we rule-out potential moral confounds that may arise when we inquire as to why the car's breaks have failed or why there are pedestrians on the highway, then we have created a realistic use case that justifies my findings. In this use case (and others like it), self-driving cars ought to respond by randomly and non-deterministically selecting who should be sacrificed. By randomly selecting any one sacrifice the self-driving car

would be saving many more people, while simultaneously recognising the equal moral status between the participants in this tragic choice scenario.

5 Conclusion

In summary, in this paper, I sought to argue how autonomous vehicles ought to react in one specific catastrophic scenario. One where an autonomous vehicle must decide who to sacrifice in a case where many can be saved if any one person can be sacrificed. While I acknowledge that the limitations I placed on this catastrophic scenario make it trivially unlikely, I hope to have advanced a plausible resolution to this specific catastrophic scenario that would bring us closer to reaching a comprehensive description of what the moral composition of autonomous vehicles ought to be.

The view that I advanced in this paper is that *in addition to whatever else the programming of autonomous cars ought to be,* autonomous vehicles ought to randomly and non-deterministically select who to sacrifice in a catastrophic scenario where any one person can be sacrificed for the sake of many more people. I argued that this view ought to hold even in cases where relatively younger, healthier or wealthier people stand to be sacrificed or in cases where owners or passengers of autonomous vehicles stand to be sacrificed.

I argued for this view in three parts. In the first part, I advanced and defended the respect-based account of moral status from the interest-based account of moral status. Following this in the second part, I argued that because the respect-based account with its threshold concept of moral status also implies a threshold concept of inviolability then people involved in a catastrophic scenario ought to face an equal chance of being sacrificed. I advanced the view that this equality of inviolability is best expressed through autonomous vehicles being programmed to randomly and non-deterministically select who to sacrifice. In the third part, I defended the random nondeterministic programming of autonomous vehicles against contra views that might hold that younger, healthier, wealthier people ought to be deprioritized from selection. I argued that if the deprioritisation of healthier, younger or wealthier people is motivated in terms of the interest-based account of variable moral considerability, then I have already giving good reasons to reject this view in favour of the respect-based account of moral status with its threshold concept of inviolability. Following this, I considered the view that may hold passengers and owners of autonomous cars should be spared from being selected for sacrifice. I argued that sparing one person over others in a catastrophic scenario in which either one could be sacrificed would be a rejection the respect-based account with its threshold concept of inviolability that we ought to endorse. The obligation of autonomous vehicles to be programmed to randomly select who to sacrifice in catastrophic scenarios (where any one person can be sacrificed to save many other people) holds even in cases where younger, healthier or wealthier people stand to be sacrificed or in cases where owners or passengers of autonomous vehicles stand to be sacrificed.

If endorsed, the programming of autonomous vehicles will need to be effectively monitored, regulated and mandatorily enforced because there is a clear incentive to purchase or reprogram an autonomous vehicle to programming that protects the owner or passenger of said car at all costs.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest I have no relevant financial or non-financial interests to disclose.

Ethical standards I declare as the sole author that I have no competing interests. Informed consent was neither obtained nor required as this research did not involve the use of human and/or animal participants or samples.

References

1. Thrun, S.: Toward robotic cars. Commun. ACM **53**, 99–106 (2010). https://doi.org/10.1145/1721654.1721679
2. Taeihagh, A., Lim, H.S.M.: Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. Transp. Rev. **39**, 103–128 (2019). https://doi.org/10.1080/01441647.2018.1494640
3. Hu, J., Bhowmick, P., Arvin, F., Lanzon, A., Lennox, B.: Cooperative control of heterogeneous connected vehicle platoons: an adaptive leader-following approach. IEEE Robot. Automation Lett. **5**, 977–984 (2020). https://doi.org/10.1109/LRA.2020.2966412
4. Gehrig, S.K., Stein, F.J., Dead reckoning and cartography using stereo vision for an autonomous car. In *Proceedings 1999 IEEE/RSJ international conference on intelligent robots and systems. Human and environment friendly robots with high intelligence and emotional quotients* (Cat. No.99CH36289), 3:1507–1512. IEEE. https://doi.org/10.1109/IROS.1999.811692.
5. Meyer, G., Dokic, J., Müller, B. (2015). Elements of a European roadmap on smart systems for automated driving. https://doi.org/10.1007/978-3-319-19078-5_13.
6. Lim, H. S., M., Taeihagh, A. (2019) Algorithmic decision-making in avs: understanding ethical and technical concerns for smart

cities. *Sustainability* 11: 5791. https://doi.org/10.3390/su11205791
7. SAE international. 2016. Automated driving levels of driving automation are defined in new sae international standard J3016. *SAE international*.
8. Rojas-Rueda, D., Nieuwenhuijsen, M.J., Khreis, H., Frumkin, H.: Autonomous vehicles and public health. Annu. Rev. Publ. Health (2019). https://doi.org/10.1146/annurev-publhealth-040119-094035
9. Larson, W., Zhao, W.: Self-driving cars and the city: Effects on sprawl, energy consumption, and housing affordability. Reg. Sci. Urban Econ. (2020). https://doi.org/10.1016/j.regsciurbeco.2019.103484
10. Bostrom, N.: Human genetic enhancements: a transhumanist perspective. Journal of Value Inquiry (2003). https://doi.org/10.1023/B:INQU.0000019037.67783.d5
11. Hevelke, A., Nida-Rümelin, J.: Responsibility for crashes of autonomous vehicles: an ethical analysis. Sci. Eng. Ethics (2015). https://doi.org/10.1007/s11948-014-9565-5
12. Thomson, J.J.: The trolley problem. Yale Law J. (1985). https://doi.org/10.2307/796133
13. Cartwright, W.: Killing and letting die: a defensible distinction. Br. Med. Bull. (1996). https://doi.org/10.1093/oxfordjournals.bmb.a011550
14. Himmelreich, J.: Never mind the trolley: the ethics of autonomous vehicles in mundane situations. Ethical Theory Moral Pract **21**, 669–684 (2018). https://doi.org/10.1007/s10677-018-9896-4
15. Tannenbaum, J., Agnieszka T., Tannenbaum, J. (2018) The Grounds of Moral Status. *Stanford Encyclopedia of Philosophy*.
16. Buchanan, A.: Moral status and Human enhancement. Philos. Public Aff. (2009). https://doi.org/10.1111/j.1088-4963.2009.01166.x
17. McMahan, J.: The ethics of killing: problems at the margins of life. Oxford University Press, Oxford (2002)
18. Darwall, S.: The second person standpoint: morality, respect, and accountability. Harvard University Press, Cambridge, Massachusetts (2006)
19. Scanlon, T.: What we owe to each other. Belknap Press, Cambridge, Massachusetts (2000)
20. Buchanan, A.: Justice, legitimacy, and self-determination. Justice, Legitimacy, and Self-Determination. (2004). https://doi.org/10.1093/0198295359.001.0001
21. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., Rahwan, I.: The moral machine experiment. Nature (2018). https://doi.org/10.1038/s41586-018-0637-6