Ethics of Self-driving Cars: A Naturalistic Approach

Selene Arfini[1] · Davide Spinelli[2] · Daniele Chiffi[3]

Abstract

The potential development of self-driving cars (also known as autonomous vehicles or AVs – particularly Level 5 AVs) has called the attention of different interested parties. Yet, there are still only a few relevant international regulations on them, no emergency patterns accepted by communities and Original Equipment Manufacturers (OEMs), and no publicly accepted solutions to some of their pending ethical problems. Thus, this paper aims to provide some possible answers to these moral and practical dilemmas. In particular, we focus on what AVs should do in no-win scenarios and on who should be held responsible for these types of decisions. A naturalistic perspective on ethics informs our proposal, which, we argue, could represent a pragmatic and realistic solution to the regulation of AVs. We discuss the proposals already set out in the current literature regarding both policy-making strategies and theoretical accounts. In fact, we consider and reject descriptive approaches to the problem as well as the option of using either a strict deontological view or a solely utilitarian one to set AVs' ethical choices. Instead, to provide concrete answers to AVs' ethical problems, we examine three hierarchical levels of decision-making processes: country-wide regulations, OEM policies, and buyers' moral attitudes. By appropriately distributing ethical decisions and considering their practical implications, we maintain that our proposal based on ethical naturalism recognizes the importance of all stakeholders and allows the most able of them to take actions (the OEMs and buyers) to reflect on the moral leeway and weight of their options.

Keywords Autonomous Vehicles · Trolley Problem · Self-driving cars · Ethical naturalism · Responsibility · Policy

---

Extended author information available on the last page of the article

🙎 Springer

1 Introduction

The potential development of self-driving cars, also known as Level 5 Autonomous Vehicles (henceforth AVs for short; BMVI 2017), may face different theoretical and practical limitations. The most recent type of self-driving car, the Level 4, still needs a human driver to request control of the vehicle in sporadic emergency cases. Waymo, a Google self-driving car, is a Level 4 Autonomous Vehicle capable of traveling more than 32 million kilometres under many different weather conditions (Waymo 2018, 2020). Of course, safety issues and technical problems are the main concerns of today's testing. In addition, ethical issues affect the concrete possibility of regulating AVs. For instance, let us consider the following scenario in which an AV should decide which action to take among some well-defined alternatives:

> An AV is driving on a high-speed road when a huge beam falls in front of it. The AV has three possibilities: go left and kill a motorcyclist without a helmet, go right and probably kill a motorcyclist with a protective helmet, go straight against the beam and kill the passenger. What should the AV do?[1]

The example above is a possible adaptation of the "Trolley Problem" (or "Trolley Dilemma"), a thought experiment ideated in moral philosophy and now used by many authors (Lin, 2016; Bonnefon et al., 2019) to map the scenarios in which AVs will find themselves in a possible future. In this specific case, every AV choice likely leads to killing at least one human being. So, which is the best decision the AV should be programmed to take? Who should be held responsible for the consequences of that decision? Providing a pragmatical solution to these pending questions should be a priority before the broad commercialization of AVs.

Yet, there is still no international public regulation on AVs and there are no behavioural emergency patterns that have been publicly accepted by both communities and OEMs (Original Equipment Manufacturers). The following documents attest to how much things are improving from an ethical perspective: the German ethics commission report (BMVI 2017), the Dutch White Paper on Ethics of Self-driving Cars (Santoni de Sio, 2016), and, more recently, the Horizon 2020 Commission Expert Group (2020), which provides 20 recommendations concerning road safety, privacy, fairness, explainability, and responsibility related to AVs. Still, suppose an OEM manages to build AVs. In that case, it will not know in advance how to specifically set them up for extreme circumstances such as the no-win scenarios (inevitable crashes) or more common traffic-related situations, thus risking legal problems and public image damages. Instead, what we propose is to involve all the stakeholders into the ethical discussion, giving the OEMs a way to set AVs in advance, the legislators the possibility to enforce limitations and buyers the chance to get involved in the process of deliberation.

Indeed, before getting too specific on our proposal's details, we need to say that we aim to discuss some key aspects associated with AVs regulation. With this intent,

---

[1] This is a no-win scenario, i.e. a situation in which an agent has choices, but no choice leads to a net gain. In other terms, the AV is going to face an unavoidable accident.

we want to distance our approach from previous literature that tried to deal with AVs' ethical problems relying solely on either a deontological (see, e.g., Powers 2006; Coca-Vila 2018) or a utilitarian perspective (Karnouskos, 2020). Indeed, we hold that a theoretical account based on either one of those moral theories cannot provide the proper ground for the ethical assessment and regulation of AVs since, as we will show, it would face too many theoretical and practical limitations. Still, we do not imply that we should simply renounce the normative dimension of ethical theories and adopt merely descriptive or social-oriented approaches to AVs' ethics like the Moral Machine Experiment (Awad et al., 2018; Noothigattu et al., 2018). Yet, we maintain that ethical theories in applied ethics cannot be blind to concrete human behaviour.

Therefore, *the solution advanced in this paper is to ground a new promising normative theory on ethical naturalism, which, unlike other moral perspectives, may provide possible guidance to the policy implementation of AVs on the streets since it is susceptible to social utilities, contextual necessities, and pragmatic adjustments but still normative-laden.* The specific form of ethical naturalism that we endorse is mainly linked to new developments in evolutive and cognitive sciences. Indeed, it is: "(1) an account of moral normativity that roots normativity in nature, where the content of nature's ontology is (provisionally) provided by the methodological canons of the natural sciences, and (2) an account of our capacity to grasp and accede to these norms that is rooted in the best theoretical frameworks that the mind sciences have to offer" (Casebeer, 2003, p. 12). We also need to say that, so far, there are just a few papers which aims at investigating and providing a normative theory based on ethical naturalism in the vast literature on AVs (a notable example is Dubljević 2020): so, we have shaped our paper to support the use of this emerging theoretical position in the framework of AI-based technologies. As we will point out later on (especially on sub-section 3.1), indeed, both science and engineering have a key role in the genesis of an ethical perspective on AVs within this account. Moreover, the regulation proposal presented in Sect. 3 could be used to cover many phases of AVs development, from the design of the policies to the after-sale evaluation, and tackling the responsibilities at each level, from owners' to national authorities'.

Among various ethical concerns that we will address in connection to the near presence of AVs on our streets, in this paper we will discuss the ethical and regulatory problems that relate to the behaviour of AVs in unavoidable accidents but also in other contexts, both in terms of preventive strategies and in post-hoc responsibility. We will suggest a way to approach these kinds of problems, which involves all the shareholders in the decision-making processes on AVs regulation.

The main idea behind the potential development to Level 5 is that AVs could drive better than humans in standard situations and emergencies (Goodall, 2014a; Boudette & Isaac, 2016). In this case, the user would have no duty of paying attention to traffic and no way of intervening to avoid a crash. Moreover, according to an optimistic view on Level 5 AVs, it is presumed that AVs' commercialization would increase the safety of the streets.[2] Every year, there are around 1.2 million deaths in car accidents

---

[2] We are aware that the assumption that the commercialization of AVs would increase and even maximise the safety of the streets is controversial in current debates on AVs production. It also becomes more

world-wide (Gogoll & Müller, 2016), and it was estimated that 90% of them are caused by human errors (Bonnefon et al., 2016, 2019). So, according to this view[3], many of these lives could be saved if AVs would be adopted. Still, such an optimistic view faces enormous practical and theoretical limitations, as we will show.

Section 2 will discuss the propositions already set out in the current literature regarding both policy-making strategies and theoretical accounts. In particular, we will reject descriptive approaches to the problem (such as the one presented by the authors of the "Moral Machine Experiment" - Awad et al., 2018; Noothigattu et al., 2018) as well as the option of using either a strict deontological view or a solely utilitarian one to set AVs' ethical choices. In Sect. 3, we will discuss our proposal's details by examining three hierarchical levels of decision-making processes: country-wide regulations, OEM policies (based on ethical, social, and technological constraints), and buyers' moral attitudes. Finally, in Sect. 4, we will further maintain that, by appropriately distributing ethical decisions and considering their practical implications, our proposal recognizes the importance of all stakeholders and allows the most able of them to take actions (the OEMs and buyers) to reflect on the moral leeway and weight of their options.

2 The Trolley Problem and Real-Life Scenarios

Applied ethics deals with the pragmatic effects of using moral reasoning, theories, and considerations into real-life contexts. In fact, when we face a problem that pertains to applied ethics we often refer to both theories of normative ethics (which aims at establishing normative standards and values) and metaethics (which studies the nature and justification of ethical languages and theories). The discussion regarding the commercialization of AVs is a paradigmatic topic of applied ethics, which in the last few years prompted extensive debates about: (1) which ethical theory should be adopted to answer a wide range of predictable moral dilemmas; (2) which metaethical tools and judgments should be used to choose a sole or dominant ethical theory; and (3) which pragmatic implications in adopting one theory or another should be held important enough to change the dominant ethical theory. In this section we will

---

complicated if layers of complexity are added: (i) if we focus on a scenario where there are mixed roads (with both AVs and regular vehicles), (ii) if we focus on the development from the commercialization of not-fully automated vehicles to fully AVs, (iii) if we reflect on a scenario in which AVs are allowed only on streets separated from regular traffic, and so on. We do not intend to engage with this controversy because we start from the assumption that even if it would be possible to prove beyond any reasonable doubt that the commercialization of AVs would dramatically increase the safety on the streets, some ethical issues that, for example, depend on unavoidable crashes, will still need to be considered and discussed. Moreover, increasing traffic safety is not the only reason why the development of AVs is seen as ethically desirable: for example, arguments are raised on the fact that it would increase the mobility and autonomy of people that now cannot drive (for medical issues and disabilities). Thus, in this paper, we start from the assumption that developing AVs can be desirable so it is useful to address some of its still unsolved moral dilemmas.

[3] The holders of this optimistic approach also assume that, at least for a long period of time, AVs and human-driven cars will coexist on the same streets, so we will have to deal with ethical issues related to having mixed roads.

proceed by critically discussing the current literature on all three of these areas of research.

We will look, once again, at the Trolley Problem, which is one of the most used metaethical tools in the discussion of moral dilemmas related to the commercialization of AVs.[4] Here is the original set up for the thought experiment:

> […] it may rather be supposed that he is the driver of a runaway tram which he can only steer from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed. […] The question is why we should say, without hesitation, that the driver should steer for the less occupied track, while most of us would be appalled at the idea that the innocent man could be framed. (Foot, 1967, p. 8)

In the original example the driver has only two options: let the tram kill five people by not steering the vehicle (so by inaction) or kill one person by steering. Since we know that the vehicle will in any case kill at least one person, if our ethical rule is to minimize the damage and maximize the implicated utility, choosing the option that ends up killing the least amount of people would naturally be the most ethical choice. Conversely, if one of our ethical rules denies the possibility to justify killing in any case, we should choose to not steer the vehicle and thus provoke the death of five people only by inaction. Those were the two main choices that Philippa Foot considered in the original take.[5]

While the choice between a utilitarian approach (which justifies the maximization of the utility and the minimization of the damage for the greatest numer of people) and a deontological one (which justifies the adherence to a finite set of rules that divide absolute duties and prohibitions) remains relevant in the current literature (Bonnefon et al., 2016; Gogoll & Müller, 2016), some new issues arise in adapting the Trolley Problem to deal with moral dilemmas concerning AVs – for example, what in the next section we will refer to as the "Chicken Problem" (Lin 2013), i.e. a situation in which a solely utilitarian approach could cost human lifes if implemented in mixed roads with human-driven cars.

We can generally describe thought experiments as heuristic tools that help us refine our pre-theoretical intuitions regarding a specific situation, which would otherwise be difficult to deal with using regular forms of argumentation (Dennett, 2013; Arfini et al., 2019). Thought experiments may help us imagine new scenarios but also evaluate existing ones in different ways. In this sense, as useful as it may be in a theoretical

---

[4] Such dilemma was originally created by Foot (1967) as a theoretical mean to analyse the doctrine of the double effect in a time in which the concept of AVs was, at best, a sci-fi dream. The doctrine of the double effect argues that it should always be considered wrong to perform a bad act intentionally in order to bring about good consequences; however, it should be sometimes permissible to perform a good act even if it will bring about bad consequences. This argument was first credited to Thomas Aquinas in the *Summa Theologica* (1988, II-II, Qu. 64, Art.7).

[5] But why is the Trolley Problem still one of the most discussed models to map the possible scenarios in which AVs may likely be involved in the future? One of the reasons advanced in the literatrure is that, since the Trolley Problem is a thought experiment, it is designed to trigger our moral intuitions (Dennett, 2013; Arfini et al., 2019) – an argument we briefly discuss at the end of this section.

debate about agents' moral intuitions, the assumption that the drivers have certain, complete and detailed information about the consequences of their actions (which means that every variable should be certain and fixed in advance) is not a realistic depiction of real-life situations. As a matter of fact, the type of AI technique adopted by AVs, i.e. machine learning, cannot by definition have complete certainty on any value. Thus, to better understand the behaviour of AVs, we need to assume that it would be impossible for it to map a scenario with certain and fixed information about the future implications of its choices (Goodall, 2014b; Lin, 2014a; Gogoll & Müller, 2016; Nyholm & Smids, 2016; BMVI 2017; Holstein 2018; Bonnefon et al., 2019). This way, all the dilemmatic situations concerning the behaviour of AVs will be based on predictions with varying confidentiality degrees. This is why the use of the Trolley Problem does not seem to help us in providing new insights on the risks, uncertainties, and the legal and moral responsibilities we encounter in daily traffic (Nyholm & Smids, 2016). Still, it is a good heuristic method to evaluate our moral intuitions.

2.1 The Solely Utilitarian vs. the Solely Deontological Approach

If the sole utilitarian approach is used, AVs would need to behave in such a way as to maximize social benefits for the greatest number of people and, possibly, to minimize global harm or damage. This approach would lead the AV in the no-win scenario of our previous example to hit the motorcyclist with the helmet: this is the only decision which could save a life, since in any of the other two options a person would very likely die.

Such choice might not seem problematic if we focus on the here and now of the accident (indeed, in that situation it would possibly mean that no one is killed by it), but it would become a problem if all AVs always chose to hit the most protected person on the road in any type of unavoidable accident. Indeed, even if that would happen only in rare emergency situations, this setting (if universally adopted) would result, over time, in penalizing responsible people who invest money in their own safety, since they would be more easily targeted. Hence, adopting this approach for all AVs may encourage people not to use safety measures in order to avoid becoming targets of other AVs (Lin, 2014b, 2016). Moreover, if a solely utilitarian approach is used, different scenarios may arise in which the best choice AVs can take involves the sacrifice of their passengers. An example is the so-called "Chicken Problem", i.e. a situation in which, on a hypothetical bridge, two human-driven cars could play chicken (Lin 2013) purposely occupying all road lanes, as they know that an utilitarian AV is going to swerve down the bridge to avoid collision with the other cars; such situation would result in killing all passengers in the AV because of a strategic move of the other cars. Such a rather unfortunate utilitarian feature may have the inconvenient consequence of making AVs less appealing on the market if not modified, regardless of the fact that AVs could be considered (at least ideally) much safer than regular cars. All the more so when we take into account that an AV could transport not only the car owners, but also their relatives, children, and so forth.

An alternative to these two problematic features of solely utilitarian AVs is to adopt a deontological approach. However, also this choice ends up creating problems if applied indiscriminately on all vehicles.

The first alternative we could consider is, for instance, the one that opposes the idea that AVs should try to hit the most protected person on the road to avoid the worst consequences at the time of the accident. A simple deontological approach suggests setting the vehicles not to discriminate between people who could be hit in no-win scenarios. Here, the problem (taken to its extreme consequences) lies in the following fact: if consequences are not evaluated, then AVs would metaphorically be tossing a coin to establish who could be harmed in such a terrible scenario. This feature would not have repercussions on the future adoption of safety measures; yet, it would not make AVs a better option with respect to human drivers especially in difficult situations, which should instead be one of the goals of AVs. Moreover, it would possibly create the same problematic situation of having to sacrifice the passenger if necessary[6].

An alternative that can be established if a solely deontological approach is adopted would be to set AVs to always protect their passengers, like some brands such as Mercedes already stated (Taylor, 2016). Unfortunately, that would have other unwanted pragmatic consequences: first and foremost, the fact that in no-win scenarios AVs would systematically target other vehicles and people, no matter how many victims they would take, to protect their passengers. This would possibly create a NIMBY (Not in My Back Yard) effect (Tamburrini, 2020): communities (if not regions) would refuse to let AVs in their streets and that, in turn, would result in delaying innovation in the field.

Still, one may imagine further deontological alternatives. For instance, something like "do not harm people who are respectful of the law". However, if a subject S commits an extremely mild infraction to a given traffic law (possibly unrelated to the accident), this cannot be *per se* a reason to blame S[7]. Another deontological alternative may be "those who impose risk should pay for the consequences". But this deontological claim can be defeated as well: all regular cars, pedestrians, AVs, etc. impose risk for the simple fact of being out on the streets. And if we cannot apply any consequentialist methodology for computing and balancing risk and benefits (e.g. a cost-benefit analysis[8]), then this last deontological alternative cannot be meaningfully implemented.

More generally, there is a more essential problem if we consider the possibility of solely adopting the deontological approach to program the behaviour of AVs. AVs would need to adhere to a finite set of principles and rules. Unfortunately, it is impossible to apply a specific set of rules for all the possible cases (as shown in Goodall 2014a, b; Holstein et al., 2018), since there will always be situations in which it will be at best impractical to respect all the rules (Gerdes & Thornton, 2016).

---

[6] We are not arguing that the AVs' setting should never allow the sacrifice of the passenger. We maintain that, by not letting the buyers choose and accept this possibility as one of the options, the innovations of AV technologies could be stopped.

[7] It is worth noting that a revision of traffic rules will be necessary once AVs will be on the streets.

[8] On a philosophical assessment of the methodology of cost-benefit and multi-criteria decision analysis, see (Hansson, 2007a, Dubljevic et al. 2021).

2.2 A Descriptive Approach: The Ethical Limitations of the Moral Machine Experiment

To respond to these fundamental problems of classical normative theories in dealing with the implications of AVs' ethics, descriptive approaches have been advanced in recent years, which study people's behavioral characteristics, perceptions, and attitudes related to AVs (Nastjuk et al., 2020) in order to reshape the ethical and normative research according to them. One of them, the "Moral Machine Experiment" has been recently proposed by an interdisciplinary group of researchers (Awad et al., 2018; Noothigattu et al., 2018) and represents a paradigmatic case of descriptive approach towards realistic scenarios. In the Moral Machine Experiment (henceforth MME) public choices made in online settings involving AVs have been expressed by collecting 40 million votes. However, this approach – as well as other solutions based on the analysis of human behaviour (Goodall, 2014a, b) – shows many critical aspects.

First of all, there are no eligibility criteria to vote. For instance, participants in the MME do not belong to a randomized group of people, and most of the voters were young and tech-savvy. This fact explains why their findings may hardly extend to the more general population to form a policy without a proper stratification of the data. Moreover, the proposed scenarios in MME handle only a fixed number of situations that may involve AVs on the streets. This is an incomplete list since many plausible scenarios are not presented and their selection, if done without a precise scientific rationale, may lead to inconclusive results. Furthermore, the participants to MME tend to discriminate among people on the basis of various characteristics (age, sex, physical shape, perceived wealth, etc.): this means that people choose which route hypothetical AVs should take not by considering *which alternative* will result in killing less people but *which people* they could end up killing. Still, even if most participants (in a randomized group) would hypothetically end up preferring one outcome over another in a given scenario, the decision of setting up AVs to discriminate between possible targets following the preferred choices of the majority would be unarguably considered unethical (Goodall, 2014a, b; IEEE 2014; BMVI 2017).

Given all these problematic issues, it is not easy to provide a descriptive framework that is both ethically sound and pragmatically feasible to be implemented in AVs. In the next section we will illustrate our approach based on a naturalized perspective on ethics as well as on a pragmatic take on responsibility in case of accidents.

3 A Mixed Approach: A Realistic Proposal Informed by Philosophical Reflection

As we argued, solely adopting either one of the two main theories of normative ethics – utilitarianism and deontology – would not solve many of the problems that ethically setting AVs brings to the fore. Moreover, we have shown that a descriptive approach toward AV ethics can also face many problems, even if it may convey some realistic facets. That is why in the literature another category of perspectives has begun to emerge in the last decade, which mixes together different normatively

oriented approaches without imposing a deontological or consequentialist view to solve moral dilemmas. These new perspectives offer metaethical methodologies to approach applied ethics issues, not always landing in a need for a case-by-case scenario. Just to mention a few of these perspectives: the method applied in the context of AVs by Dubljevic et al. (2021, p. 2) based on the *multi-criteria decision analysis*, which is a way to study potential harms and risks of a given technology by considering "the multi-faceted impacts of technological change;" the *function-based working approach*, developed by Fossa et al. (2022, p. 1), which aims at developing methodological tools to support "the exercise of moral judgment aimed at aligning AV design to the EU normative framework"; the theoretical approach defended by Tamburrini (2020) and Fossa & Tamburrini (2022, p. 81), which view moral dilemmas as "pointers to the need of striking trade-offs between values at stake". Our perspective, then, falls directly into this mixed approaches category. Indeed, in order to find an ethical and pragmatic compromise to set AVs, we will try to fruitfully combine both ethical approaches, whilst keeping an eye on our few realistic goals (finding a way to distribute the ethical decision-making processes to involve all stakeholders and allow them to reflect and take action considering the weight of their options), which are related to the hypotheses we took as our theoretical ground in the introduction to this paper.

3.1 Philosophical Justification

To justify our mixed ethical approach from a metaethical point of view, we can consider a naturalistic perspective on morality. Such perspective on ethics implies that the moral choices and norms that a particular community adopts emerge from social utility, contextual necessities, and pragmatic adjustments between possibilities and goals of the agents who are immerse in changing socio-cultural environments (Casabeer 2003; Magnani 2011; Fitzpatrick, 2016). Ethical naturalism does not amount to relativism since it is based on results from different scientific theories and perspectives (such as the ones that inform evolutionary biology, psychology, and cognitive science) on the origins, nature, and development of morality. Hence, the ontological presuppositions of ethical naturalism change when (and as soon as) new scientific ideas arise and are proven in the scientific community (it basically changes and is updated with the emergence of new evidence and insight). In the naturalistic perspective, the socio-cultural environment in which human agents are embedded shape them morally, while the normative ethics adopted by a community always result from a compromise between the ethical and pragmatic goals, possibilities, and constraints of its agents.

Based on a naturalized ethical approach we may inform a pragmatic perspective on how to regulate AVs in an effective way. The main idea is to suggest potential regulations for AVs inspired by ethical reflection. From a regulatory perspective this implies that, for example, it is necessary to create a country-wide regulation for AVs. That regulation should be, however, quite uniform in large regions of the world (e.g. continents) in order to avoid great differences among neighbouring nations. In line with these general (and possibly uniform) regulations, every OEM would develop different policies and codes for their AVs, basing them on ethical, social, and technological constraints. Then, before the purchase of an AV, each buyer could test per-

sonal ethical attitudes through an interactive experience - as realistic as possible - in a certified setting (e.g. Virtual Reality, Faulhaber et al., 2018).

In many countries, in case of an accident, the responsibility lies on the drivers, whether or not they are the car's legal owners (apart from rare cases when the owner's negligence is more relevant and proven). With a self-driving vehicle, the "driver", intended as the person sitting at the driver's seat in the car, is nothing more than a passenger and the "actual driver", the entity who leads the AV, is an AI program together with the complex systems of sensors and actuators, which cannot be held responsible *per se* in case of an accident. Thus, the liability would fall on the car's owner, who chose the policy, as well as on the OEM, who provided it (as cited in Rule 1 of Borenstein et al., 2017, p. 394). Nonetheless, some forms of partial agency may be attributed to AVs and this may be characterized by a human-technology interface, for instance when an AV has to achieve a specific human goal (Nyholm, 2018a, b). This means that a complex interaction between human and machine is unavoidable.

3.2 The Analytical Defence of Our Proposal

Let us discuss analytically our proposal. The decision to have country-wide regulations derives straight from the naturalized perspective on ethics which acknowledges that many countries have different legislations and cultural perspectives (Awad et al., 2018). In some countries, age-based discriminations of the targets would be illegal (as delineated, for example, in the German ethics commission report, BMVI 2017), while in others age could be considered as a feature to safeguard children and other high-risk categories in case of an accident. Even if we acknowledge that there may exist country-based differences of culture and moral attitudes, we have to admit that it would also be quite impractical to have diverging policies on AVs in different countries of the same region of the world. Therefore, the creation of supranational organizations is required to harmonize, not replace, the national standards of regulations, producing general declarations and codes of conduct. As a way of example, let us imagine a European Safety Council for AVs aimed at reducing the number of accidents involving AVs in Europe, whose members are experts proposed by national states, European institutions, and representatives of all stakeholders.

Subsequently, every OEM could develop different policies, as a set or as parameterized functions, in order to give buyers the possibility to choose their own AV insurance policy coverage based on a critical reflection on (and a selection of) a restricted range of alternatives, coherently with regulatory and ethical constraints. Giving this possibility is important because we argue that users have the right to decide how their AVs should behave consistently with the alternatives that are morally admissible (Lin 2013, 2016).

Our perspective "avoids paternalistic impositions of a policy" (Tamburrini, 2020) and allows users to reflect on the moral leeway and weight of their actions, whilst possibly letting them take part as individuals in the discussion on moral responsibility related issues. A collective responsibility of all users in the form of "strict liability" has been also proposed (Hevelke & Nida-Rümelin, 2015); yet, we think that assigning collective responsibility among all users can be problematic in many contexts of use for AVs and not always fully justified from an ethical perspective. It is not clear

why these costs should rely completely on owners and users and never on OEMs. The invention of new insurance systems for AVs in order to assign liability (also based on philosophical reflections) is, in fact, one of the last recommendations of the Horizon 2020 Commission Expert Group (2020).

The pragmatic strategies associated to our proposal are the following:

i.   OEMs would be allowed in some contexts to introduce ad-hoc restrictions in order to protect their buyers (Lin, 2016). This means that OEMs should require an approval for the design of a new model of AV to an independent ethics committee composed of a group of relevant stakeholders and experts in the field. The ethics committee would ground its decisions on ethical declarations and documents promoted by international regulatory agencies.

ii. Users may profit from a robust informed consent in which the risks and benefits of this technology are fully explained and possibly accepted. It is worth noting that informed consent should not be considered as mere "terms of use" but rather as a critical reflection and acceptance by the user of the risks of a specific model of AV. Thus, OEMs have the duty to clearly inform customers, owners, and users of AVs. It can be moreover argued that pedestrians may be involved in accidents with AVs, which is why social forms of informed consent for the acceptance of AVs should be implemented for pedestrians as well.

iii. The ethical test would be carried out on buyers since they would be the owners of the AVs and, therefore, they might be held responsible for accidents, even if they are not physically in their AVs at the moment of the collision. This may be so, since the AV could have behaved just following the insurance policy chosen by the owner. There should be an interactive part of the test aimed at taking into consideration and evaluating the buyer's decisions. Then, these decisions could be analysed and reported by a trained person to the buyers. Following this protocol, buyers would be able to adjust their choices in a relaxed and insightful environment. It is worth noting that since the AVs may be used by people other than the owner, the ethical test should include hypothetical scenarios of this kind, which may pose intricate problems regarding the insurance policies for AVs. At the same time, the test and report would need to be adjusted also to consider situations of shared ownership so as to distribute to all owners both access to the ethical decision-making and responsibility.[9] However, the difference between realistic rapid decisions and unrealistic slow decisions can make a huge difference in evaluating a user's behaviour, which is why this factor should also be included in the test and report[10].

Similar to what can be observed with regular car accidents, the owners and the OEMs would be able to buy insurances with various degrees of coverage. The difference is

---

[9] We reckon that another possibility should be taken into account: the case in which society will not allow individual ownlership of AVs and only commercial fleet operators could own the AVs. This scenario is amply discussed by Dubljević et al., (2021). We believe that the consideration of our approach in the fleet ownership scenario deserves to be apprached in depth in a future analysis.

[10] Recent results show that while the utilitarian choice had marginal prevalence when virtual choices are made in non-emergency situations, it dominated in emergency situations (Lucifora et al., 2020).

that the information regarding the policy chosen would be a characteristic evaluated for the computation of the insurance premium. For example, if the owner chose a balanced but still slightly 'selfish' policy, this fact would slightly increase his or her insurance costs as a direct consequence[11]. The OEM may be (fully or partly) held responsible for an accident, depending on the specific situation. This could be a strong incentive for OEMs to constantly improve the performance of AVs to avoid emergencies. However, this also depends on whether AVs will be equipped with autonomous maintenance solutions. If maintenance is not autonomous, then the owner failing to secure maintenance procedures for his or her AVs might be considered responsible in those crashes due to his or her negligence. Still, regulators must impose periodical maintenance procedures as is done with regular cars.

3.3 Expected Consequences

In our framework, OEMs can be held responsible from an ethical and legal perspective for a great amount of crashes, which is coherent with some recent views stated by notable OEMs. For instance, AUDI and Volvo have already declared their willingness to be held responsible for crashes involving AVs (Atiyeh 2015; Maric, 2017). In virtue of these considerations, OEMs could also decide to retaliate on the producers of the components that would cause a large number of accidents, encouraging, in turn, their improvement. Holding the OEMs (at least partially) responsible for the behaviour of its AVs would indeed trigger a chain of developing improvements regarding the safety of AVs in order to avoid penalties and public image damage. Owners will be considered responsible of the crashes with AVs in case of negligence in their maintenance (if this is not automatic).

From an ethical point of view, all these measures would provide a good compromise between a fully utilitarian or a fully deontological approach in a way that is also sensitive to social preferences, thus improving the social acceptability of AVs intended as a type of potentially disruptive innovation[12]. In sum, our approach will

i.   contribute to acknowledging the responsibilities of OEMs and owners, providing them with ethical guidelines.
ii.   avoid the occurrence of the Chicken problem, since there would be many different policies, making it impossible to figure out which one is implemented by a car just by looking at it from the outside.
iii.   entail that there would be no need of a system in which all costs of the crashes of AVs are shared by all owners/users of AVs in the form of "strict liability".
iv.   imply that it would be mandatory for OEMs to share all relevant information to the competent authorities in case of accidents, in order to reconstruct exactly what happened and how the AV reached a specific conclusion (Hevelke

---

[11] To avoid the fact that only rich people may partially increase the policy associated to their AVs in a balanced but still slightly 'selfish' way, some clear limitations should be imposed, and the sharing of AVs should be promoted, too.

[12] On the connection between moral uncertainty and disruptive innovation, see (Nickel, 2020) and (Nickel et al., 2022).

& Nida-Rümelin, 2015; Holstein 2018), though consistent with privacy-related constraints.

The main opposition to our realistic approach could be moved by pedestrians and bikers who would be targeted by the AV. Pedestrians and bikers should be particularly protected. We follow, in this case, the German ethics commission, which states that pedestrians and uninvolved parties should not be targeted by AVs (BMVI 2017, p. 11). It is worth noting that in our framework, only the insurance policy can be slightly modified, and not the user's ethical setting for AVs. We believe that an "Ethical Knob" (Contissa et al., 2017), i.e. a device enabling passengers to ethically customise their AVs, clearly involves many limitations since people might then potentially choose their ethics setting based on unacceptable ethical ideas and this might be particularly dangerous for pedestrians and uninvolved parties.

In order to protect them, we could imagine some forms of social experimentation to be done by society as a whole in order to cope with the rights of pedestrians. Similar to what happens in medicine with the post-market surveillance of drugs, it is imaginable to have some form of similar surveillance for AVs[13]. This may also be achieved by involving all relevant stakeholders in traffic regulation and by means of the introduction of AVs into society as a social experiment (Van de Poel, 2016); an approach that seems to be coherent with our naturalistic ethical perspective and is particularly welcome in order to cope with the fundamental uncertainty associated to the behaviours and movements of individuals in a society[14]. This is especially true in those scenarios in which regular cars, AVs and pedestrians use the same roads (which, we assumed, will be a common situation for a while, at first). Owners and users may take decisions regarding some ethical aspects of their AVs, but their choices will be constrained by competent authorities. Finally, it is worth noting that rejecting the possibility to adopt a balanced but still slightly 'selfish' policy may lead to a lower number of AVs on the market regardless of their expected good performances on the roads; in turn, the total harm brought by human-driven cars to pedestrians might never decrease.

---

[13] London & Danks (2018) suggested the possibility that AV regulation may resemble the regulation of drugs and other medical interventions. Even if we are sympathetic with this view, we still think that engineering sciences have already many stages and resources to address value-based considerations. For instance, during the design process we have the following 'phases' (not occurring necessarily in this order): problem analysis, conceptual design, simulation, decision stage, detail design, prototype development and testing, which may be implemented in order to deal with the ethical challenges of AVs. There are also methods often used in engineering to deal with value conflict such as cost-benefit analysis, multiple criteria decision analysis, value-sensitive design, just to mention some of them. A philosophical analysis of some of these methods has been recently discussed in (Hirsch Hadorn, 2021). Unlike cost-benefit analysis, which is a clear consequentialist methodology, multiple criteria decision analysis may be a proper tool to deal with mixed approaches towards regulating AVs. Moreover, in multiple criteria decision analysis, the ethical, social and environmental risks of AVs, which are usually difficult to quantify, are not expressed by means of monetary units as it happens in cost benefit analysis. Finally, multiple criteria decision analysis may help us identify objectives, formulate decision criteria and establish their weights in order to assess the reasons behind a specific ranking of alternatives.

[14] On the relation between fundamental uncertainty and moral leeway, see (Hansson, 2007b, 2011; Chiffi & Pietarinen, 2017). On the different logical levels of ignorance and uncertainty, see (Carrara et al., 2021).

4 Conclusions

In this paper, we offered a new perspective based on *ethical naturalism* that could help reframe some ethical and regulatory issues related to the setup and commercialization of AVs. We began by describing one of the most problematic situations in which a future AV could soon find itself: the no-win scenario. The typical questions related to these kinds of cases are: what choice should the AV be programmed to make? Who should be held responsible for it?

Focusing on the first question, we argued against adopting one of the traditional normative theories – utilitarianism and deontology – so as to find a viable answer that could work for any AV. Choosing either one of these theories would have negative repercussions on the use of AVs as well as on the social and safety norms of the communities in which they would drive. We also described the ineffectiveness of choosing the best alternative by using a descriptive account of consumers' ethical attitudes as proposed by the MME.

In fact, choosing one option and applying it to any AV would bring out unethical or impractical consequences. Moreover, it would deprive the OEM and the AVs' buyers of the right (and the burden) of making ethical choices, and in turn, of being held accountable for them. To bridge these gaps and to critically discuss the problems raised by unavoidable accidents, we divided up the decision-making process into three hierarchical levels.

1. We imagine country-wide general regulations (harmonized by supranational organizations) to acknowledge that different legislations and cultural perspectives could bring different forms of rules for the AVs' behaviour in various parts of the world.
2. Within the parameters fixed by country-wide regulations, OEMs would then provide different ethical policies. This requirement would make OEMs both active agents in the ethical decision-making process and partly responsible for their AVs' behaviour.
3. Buyers could choose the insurance policy that better fits their moral attitudes, thus enabling them to reflect on the moral leeway and weight of their own actions.

We devised these measures to be in line with the premises described at the beginning of the article. Undoubtedly, society will need to organize forms of social experimentation to adjust to the changes that will inevitably occur when the first AVs will appear on our streets; yet, the measures we have discussed may provide spacious elbow room not only to adapt to these changes, but also to improve hopefully AV technology as a whole.

Declarations

References

Aquinas, T. (1988). (13th c). Summa Theologica II-II, Q. 64, art. 7, "Of Killing". In R.J. Regan and W.P. Baumgarth (Eds.), *On Law, Morality, and Politics*, 226–227. Indianapolis/Cambridge: Hackett Publishing Co.,

Arfini, S., Casadio, C., & Magnani, L. (2019). Ignorance-preserving mental models. Thought experiments as abductive metaphors. *Foundations of Science*, 24, 391–409

Atiyeh, C. (2015). Volvo will take responsibility if its self-driving cars crash, *Car and Driver* October 8, 2015. Retrieved from: https://www.caranddriver.com/news/a15352720/volvo-will-take-responsibil-ity-if-its-self-driving-cars-crash/ (Accessed 22 September, 2020)

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A. … Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729), 59–64

BMVI. Bundesministerium für Verkehr und digitale Infrastruktur - Federal Ministry of Transport and Digital Infrastructure (2017). *Ethics Commission: Automated and Connected Driving*. Retrieved from https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile

Bonnefon, J., Shariff, A., & Rahwan, I. (2019). The Trolley, The Bull Bar, and Why Engineers Should Care About The Ethics of Autonomous Cars [point of view]. *Proceedings of The IEEE*, 107(3), 502–504

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

Borenstein, J., Herkert, J., & Miller, K. (2017). Self-Driving Cars and Engineering Ethics: The Need for a System Level Analysis. *Science and Engineering Ethics*, 25(2), 383–398. doi: https://doi.org/10.1007/s11948-017-0006-0

Boudette, N., & Isaac, M. (2016). Head of Fiat Chrysler Sees Self-Driving Cars in Five Years, Not 20. Retrieved 22 September 2020, from https://www.nytimes.com/2016/05/07/business/fiat-chrysler-chief-sees-self-driving-technology-in-five-years.html

Carrara, M., Chiffi, D., De Florio, C., & Pietarinen, A.-V. (2021). We don't know we don't know: Asserting ignorance. Synthese, 198, 565–580.

Casebeer, W. (2003). *Natural Ethical Facts: Evolution, Connectionism, and Moral Cognition*. Cambridge, MA: MIT Press

Chiffi, D., & Pietarinen, A. V. (2017). Fundamental uncertainty and values. *Philosophia*, *45*(3), 1027–1037

Coca-Vila, I. (2018). Self-driving cars in dilemmatic situations: An approach based on the theory of justification in criminal law. *Criminal Law and Philosophy*, *12*(1), 59–82

Contissa, G., Lagioia, F., & Sartor, G. (2017). The Ethical Knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, *25*(3), 365–378

Dennett, D. C. (2013). *Intuition Pumps and Other Tools for Thinking*. New York: WW Norton & Company

Dubljević, V. (2020). Toward implementing the ADC model of moral judgment in autonomous vehicles. *Science and Engineering Ethics*, *26*(5), 2461–2472

Dubljević, V., List, G., Milojevich, J., Ajmeri, N., Bauer, W. A., Singh, M. P. … Samandar, M. S. (2021). Toward a rational and ethical sociotechnical system of autonomous vehicles: A novel application of multi-criteria decision analysis. *Plos one*, *16*(8), e0256224

Faulhaber, A., Dittmer, A., Blind, F., Wächter, M., Timm, S., et al. (2018). Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles. *Science and Engineering Ethics*, *25*(2), 399–418. doi: https://doi.org/10.1007/s11948-018-0020-x

Fitzpatrick, W. (2016). Morality and Evolutionary Biology. In Zalta EN (Ed.) *Stanford Encyclopedia of Philosophy*. Available online at: https://plato.stanford.edu/entries/morality-biology/

Foot, P. (1967). The problem of abortion and the doctrine of the double effect, Oxford Review, *5*, 5–15

Fossa, F., Arrigoni, S., Caruso, G., Cholakkal, H. H., Dahal, P., Matteucci, M., & Cheli, F. (2022). Operationalizing the Ethics of Connected and Automated Vehicles: An Engineering Perspective. *International Journal of Technoethics*, *13*(1), 1–20

Fossa, F., & Tamburrini, G. (2022). Ethics of Autonomous Vehicles. From Moral Dilemmas to Social Trade-Offs. *Paradigmi*, *40*(1), 79–94

Gerdes, J., & Thornton, S. (2016). Implementable Ethics for Autonomous Vehicles. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomes Fahren* (pp. 87–102). Berlin, Heidelberg: Springer Vieweg

Gogoll, J., & Müller, J. (2016). Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Science And Engineering Ethics*, *23*(3), 681–700

Goodall, N. (2014a). Machine Ethics and Automated Vehicles. In G. Meyer, & S. Beiker (Eds.), *Road Vehicle Automation* (pp. 93–102). Berlin: Springer

Goodall, N. (2014b). Ethical Decision Making during Automated Vehicle Crashes. *Transportation Research Record: Journal of The Transportation Research Board*, *2424*(1), 58–65

Hansson, S. O. (2007a). Philosophical Problems in Cost-Benefit Analysis. *Economics & Philosophy*, *23*(2), 163–183

Hansson, S. O. (2007b). Hypothetical retrospection. *Ethical Theory and Moral Practice*, *10*(2), 145–157

Hansson, S. O. (2011). Coping with the unpredictable effects of future technologies. *Philosophy & Technology*, *24*(2), 137–149

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *Science and Engineering Ethics*, *21*(3), 619–630

Hirsch Hadorn, G. (2021). Which Methods Are Useful to Justify Public Policies? An Analysis of Cost–Benefit Analysis, Multi-Criteria Decision Analysis, and Non-Aggregate Indicator Systems. *Journal for General Philosophy of Science*. https://doi.org/10.1007/s10838-021-09580-4

Holstein, T., Dodig-Crnkovic, G., & Pelliccione, P. (2018). Ethical and Social Aspects of Self-Driving Cars. ArXiv, abs/1802.04103.

Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (2020). *Ethics of Connected and Automated Vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility*. 2020. Publication Office of the European Union: Luxembourg

IEEE (2014). IEEE Code of Conduct. Retrieved 12 April 2022, from https://www.ieee.org/content/dam/ieee-org/ieee/web/org/about/ieee_code_of_conduct.pdf

Karnouskos, S. (2020). The role of utilitarianism, self-safety, and technology in the acceptance of self-driving cars. *Cognition Technology & Work*, 1–9. https://doi.org/10.1007/s10111-020-00649-6

Lin, P. (2014a). Here's a Terrible Idea: Robot Cars with Adjustable Ethics Settings. Retrieved 22 September 2020, from https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/

Lin, P. (2014b). The Robot Car of Tomorrow May Just Be Programmed to Hit You. Retrieved 22 September 2020, from https://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/

Lin, P. (2016). Why Ethics Matters for Autonomous Cars. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomes Fahren* (pp. 69–85). Berlin, Heidelberg: Springer Vieweg

London, A. J., & Danks, D. (2018). Regulating Autonomous Vehicles: A Policy Proposal. In *Proceedings of the 2018 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*. doi:https://doi.org/10.1145/3278721.3278763

Lucifora, C., Grasso, G. M., Perconti, P., & Plebe, A. (2020). Moral dilemmas in self-driving cars. *Rivista Internazionale di Filosofia e Psicologia*, *11*(2), 238–250

Magnani, L. (2011). *Understanding Violence. Morality, Religion, and Violence Intertwined: a Philosophical Stance*. Berlin/Heidelberg: Springer

Maric, P. (2017). Audi to take full responsibility in event of autonomous vehicle crash, *Car Advice* September 11, 2017. Retrieved from: http://www.caradvice.com.au/582380/audi-to-take-full-responsibility-in-event-of-autonomous?vehiclecrash/ (Accessed 22 September 2020)

Nastjuk, I., Herrenkind, B., Marrone, M., Brendel, A. B., & Kolbe, L. M. (2020). What drives the acceptance of autonomous driving? An investigation of acceptance factors from an end-user's perspective. *Technological Forecasting and Social Change*, *161*, 120319

Nickel, P. J. (2020). Disruptive innovation and moral uncertainty. *Nanoethics*, *14*(3), 259–269

Nickel, P. J., Kudina, O., & van de Poel, I. (2022). Moral Uncertainty in Technomoral Change: Bridging the Explanatory Gap. *Perspectives on Science*, *30*(2), 260–283

Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. (2018). In *AAAI Conference on Artificial Intelligence*. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17052/15857

Nyholm, S. (2018a). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics*, *24*(4), 1201–1219

Nyholm, S. (2018b). The ethics of crashes with self-driving cars: A roadmap, II. *Philosophy Compass*, *13*(7), e12506

Nyholm, S., & Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice*, *19*(5), 1275–1289

Powers, T. M. (2006). Prospects for a Kantian machine. *IEEE Intelligent Systems*, *21*(4), 46–51

Santoni de Sio, F. (2016). *Ethics and Self-Driving Cars: A White Paper on Responsible Innovation in Automated Driving Systems*. Dutch Ministry of Transportation and Infrastructure Rijkswaterstaat

Tamburrini, G. (2020). *Etica delle Macchine. Dilemmi Morali per Robotica e Intelligenza Artificiale*. Torino: Carocci Editore

Taylor, M. (2016). Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians. Retrieved from: https://www.caranddriver.com/news/a15344706/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/ (Accessed 22 September, 2020)

Van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics*, *22*(3), 667–686

Waymo (2018). Waymo Safety Report on the Road to Fully Self-Driving. Retrieved 13 April 2022, from https://storage.googleapis.com/sdc-prod/v1/safety-report/Safety%20Report%202018.pdf

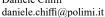Waymo (2020). Safety Report Webpage. Retrieved 13 April 2022, from https://waymo.com/safety/

Authors and Affiliations

Selene Arfini[1] · Davide Spinelli[2] · Daniele Chiffi[3]

✉ Selene Arfini
selene.arfini@unipv.it

Davide Spinelli
davide2.spinelli@mail.polimi.it

Daniele Chiffi
daniele.chiffi@polimi.it

1   University of Pavia, Piazza Botta, 6, 27100 Pavia, Italy

2   Politecnico di Milano, Milan, Italy

3   Politecnico di Milano, Milan, Italy