A White Paper Discussing Ethical Considerations for Autonomous Vehicles

C. Menon*

*University of Hertfordshire, UK, c.menon@herts.ac.uk

1 Introduction

Autonomous systems (a category which includes AVs) have been proposed for use in multiple domains, with examples including nuclear containment, defence systems, health and transport. In this paper we discuss the ethical landscape surrounding the introduction and operation of autonomous vehicles as a form of transport on the public road network. Use of autonomous vehicles in other domains (e.g. as a defence capability or a form of medical transport) is likely to impose ethical requirements which go beyond the scope of this document, and for which we refer the reader to existing literature.

The primary focus of this white paper will be on the intersection of safety and ethics for AVs. However, we note that any recommendations for the ethical introduction and operation of AVs should also consider issues such as wealth equality, manufacturing practices and environmental impact.

2 Ethical background

The "trolley problem" refers to a well-known thought experiment, in which a train / trolley is on a set of tracks which will cause it to collide with a number of people. The observer is asked whether s/he would choose to switch the train to a second set of tracks which will cause it to collide with a single person only. Amendments and extensions to the trolley problem have couched the problem in terms of an active vs passive choice as well as experimented with the relative "worth" of each person affected.

The trolley problem has a clear analogue in the case of AV behaviour, in that a situation may be encountered in which a collision with at least one group of people is inevitable. In this case, the developers responsible for the behaviour of the AV must address a trolley problem: which group(s) should the AV choose to impact? This is explored further in [2].

2.1 Systems of ethics

The trolley problem can be used to illustrate a number of different ethical systems, providing examples of how these might differ in their application to AV behaviour.

Consequentialism [3] is often considered to provide a reasonable foundation for discussion of AV ethics and behaviour. Consequentialism is an ethical theory which prioritises the outcomes: consequentialist ethics deems acts to be morally acceptable if they lead to a good outcome. This is sometimes summarised as "the end justifies the means". A consequentialist approach to AV safety would be to seek to reduce overall harm by minimising the number of people harmed; a consequentialist solution to the trolley problem would be to switch the trolley onto the section of the track with a single person. Consequentialism as an ethical theory is aligned with more general safety criteria [4] in terms of minimising harm, but does not take into account questions of risk responsibility, informed consent for acceptance of risk and calculations relating to acceptable exposure due to work.

By contrast, deontological theories of ethics prioritise acting in accordance with explicitly stated duties and rules [5]. Deontology therefore does not require the AV to consider the outcomes, but merely to act in accordance with pre-programmed rules (which may include, for example, a rule that the AV must not injure – or cause to be injured – any person). While encoding such rules is conceptually simpler than requiring the AV to perform calculations minimising harm, deontological ethics does require the identification of rules for every situation the AV may find itself in. A deontological approach to the trolley problem would be to consider whether rules exist which govern the acceptability of switching the trolley to a different track, regardless of the risk exposure to any individuals.

A third ethical imperative relevant to AVs is the concept of virtue ethics, typically presented in terms of self-sacrifice [6]. This discusses the extent to which an AV should choose to sacrifice itself and its passenger when placed in a situation in which this would reduce harm to a third party.

2.2 Extensions of the trolley problem

More generally, from a safety perspective we are concerned about the risk posed by the AV to different groups, and the ethical justification for prioritising the safety of one group over another. This extends the trolley problem to other situations in which the risk is the deciding factor. In the following examples where we refer to the decisions or choices made by the AV, this is to be understood to be the decisions and choices made by the AV system developers which result in the defined behaviour.

In [6] a case is presented whereby an AV may choose to position itself within its lane so that it is closer to a smaller car than to a truck. This decision might be justified in two ways: firstly, that this behaviour is typical of a human driver, and secondly that this reduces the risk to the AV (a collision with a small car is likely to result in less harm to the occupants of the AV than a collision with a truck). From a

safety perspective, this decision has prioritised the safety of the AV occupants – and the truck occupants – over that of the smaller car. Such a decision would need to be justified within the safety case and from an ethical perspective.

Another situation arises whereby an AV may take the opposite course; choosing to drive closer to (or in the worst case, impact) a heavier vehicle, or a vehicle with safety systems which are known to be better [6]. In this case the overall severity of an accident may be reduced, compared to an impact with a vehicle with poor safety systems. However, implementing such a decision into the behaviour of the AV represents a deliberate choice to increase the risk to drivers of certain vehicles known for their safety features. Again, this decision would need to be justified both ethically and in the safety case.

Other situations discussed in the existing literatures include the decision of an AV to sacrifice itself (place itself in the path of another vehicle to save a third party from impact) [6], as well as choosing to impact a motorcyclist wearing a helmet over one not wearing such protective devices [7].

3 Safety and ethical landscape

As we discussed in Section 1, the ethical landscape surrounding the introduction of AVs is not limited only to the trolley problem and to AV behaviour during collisions. While we do not go into detail on the ethical issues which are not directly relevant to safety (e.g. environmental impact, job loss, capability benefits, inequality of access to technology etc.), there are a number of issues which do impact indirectly on the safety considerations for AVs.

The first of these is the question of commercial forces driving early adoption of AVs. There is significant public interest in AVs, particularly around self-driving cars, and engineering companies are alert to the advantage of bringing out the "first of kind" of an AV. However, unlike the military and nuclear domains, the high-profile nature of commercial AVs can encourage thinking which views safety as a "competitive advantage". This means that known problems may not be shared for reasons of commercial interest. Furthermore, there is a potential issue with technology introductions being driven by – and dependent on – public interest. This can lead to "hyped" innovations and features being prioritised by manufacturers in an attempt to gain an increased market share from early adopters [18]. However, particularly for safety-critical systems such as AVs, the problems of assuring such features are potentially significant, and it is likely that adequate assurance will lag behind development.

In addition, there are currently no applicable standards which fully address the safety of AVs, including safety of the intended function. That is, there is a gap between the traditional automotive safety approach [8], and that required for AV operation. In some domains best practice can be explicitly appealed to where guidance and standards are incomplete, but the status of AVs as a new technology means that there is no existing best practice for these systems.

Consequently, while there is a clear economic and reputational imperative for a company to bring out the "first of kind" in autonomous vehicles, it is much less clear that such an AV could be demonstrated to be acceptably safe. There is a risk that the push to produce and market AVs can encourage "quick and dirty" practices during the development lifecycle which can have an effect on the system as released to the public. While standards do exist around ethical design of systems [9], these are relatively new and their general applicability has not been fully determined.

4 Risk transfer and consent

Perhaps the most significant issue relating to the ethics and safety of AVs is the question of risk. The risk posed by AVs cannot be assumed to be identical to the risk posed by human drivers, either in absolute terms or in terms of the distribution across different exposed groups. As we have previously discussed, there is as yet no accepted solution to the problem of providing adequate assurance for an AV, and the question of risk assessment becomes one of dealing with "unknown unknowns".

In more detail, the task of determining the differences between the risk posed by an individual AV vs a human driver – and by multiple AVs interacting with each other and human drivers – is a non-trivial problem. The answer will depend to a large extent on the technological solutions implemented: the failure modes of the AV software, the algorithmic decisions made, the efficacy of the AV hardware including sensors, and so forth.

It may be the case that introduction of AVs results in an overall temporary increase in risk – as human drivers adapt to the systems, as new failure modes emerge and as risk mitigation decisions are made – but leads to a situation where over the long-term fewer lives are lost on the road. The argument for accepting AVs in this case is founded on a claim that their introduction will result in a greater good. However, from a safety perspective, the harm that is done in the short-term must be justified. Existing standards [11] discuss the acceptability of a temporary increase in risk in return for a longer-term decrease, but it is not clear that these standards can be applied to the introduction of AVs, or that there is any exchange mechanism which can justify the acceptance of the specific deaths caused by AVs in exchange for the specific deaths caused by human drivers. A comparison may usefully be drawn here to the defence domain, where risk to operational personnel is accepted for the greater good, as it is also in commercial aviation.

However, we are interested not just in the overall risk posed by the AV, but in the distribution of that risk (and how it compares to the distribution of risk posed by a human driver). Even if the overall risk posed by an AV is lower, it may be the case that a segment of the population bears an unfair degree of this risk, i.e. there is either an absolute or a relative increase in the proportion of risk to which they are exposed. This would be the case if certain segments of the population interacted with AVs in a way that led to increased accidents

(for example, groups of pedestrians with characteristics that cause the AV sensors to miss them). This increase must, from a safety perspective, be justified. A parallel here is to the use of vaccination programs, where the overall risk due to disease is lowered, but at a cost of harming some medically-vulnerable individuals. These individuals are facing an increased degree of risk due to the vaccination program, although the overall system risk to the general public is lowered. Fluoridisation of the water supply is another relevant example.

The question of consent to risk exposure must also be considered. Given the complexity of assessing all risks posed by an AV, and the difficulty of communicating all aspects of risk management to the public, it is likely that people will be unaware of the full risk landscape around AV use. Without this information, it is impossible to say that the public has consented to the risk posed by these systems. Furthermore, even in the case where an individual may consent to a known set of risks (suppose the very unlikely situation where the passenger of an AV is in possession of full knowledge of the functional decisions and risk mitigations decisions the AV will make), other road users, pedestrians etc. have not consented to the portion of risk to which they are exposed by the decision of the passenger to use an AV.

There is also a more general risk acceptability question. Even should an AV function in exactly the same way as a human driver, it is not clear that the public will be willing to accept the same risk when it is posed by a machine as opposed to a person.

4.1 Risk inequalities

Although the primary question may be one of public acceptance of risks, it is also clear that the ownership of, and responsibility for, risks will be changed by AV introduction. When human drivers are replaced by AVs, the decisions as to what action should be taken in a collision situation are moved from a time-critical frame (just before the collision) to a frame which is not time-critical (development of the AV system and its software). This may raise the standard of ethical performance the public expects. In the case of a human driver, any decisions made in a collision situation are judged according to that environment (e.g. there is little time to choose between different options, the drivers are under stress, and – except where their actions have been negligent – are generally not considered culpable should they make the "wrong" decision [6]). However, an engineer developing the AV is not under the same pressure, and may therefore be expected to ensure that the AV reacts in a morally acceptable way, regardless of how a human driver might. In this sense the engineer is said to own the risk in a way which a human driver may not be required to. In this sense, an additional degree of risk – which must be owned by someone – can be said to have been introduced.

The impact of AVs on the wider road network is another source of additional risk. The road network can be viewed as a system of systems (SoS), with the AVs comprising one component only. The risk posed by an AV may therefore affect any portion of this network, leading to unforeseen interactions and emergent behaviour. One example of this may be an increase in traffic jams due to all AVs following the same route, as it is in the interest of no individual AV to change route. Another example may be the effect on driver norms where, for example, human drivers may customarily let other vehicles exit from a side street and the road planning is such that it presumes this type of essentially human interaction. These situations will be exacerbated in the case of AVs which make use of machine learning algorithms, where local optimisations made by these algorithms can negatively affect traffic flow, safety or efficiency of the wider network.

4.2 ALARP and AV risk

The concerns around AV introduction must also be assessed within the current legal framework of applicable standards and policies. In the UK, this means that systems must be shown to be ALARP [4]. The ALARP framework deals solely with safety, and has no mechanism for trading an increase in risk for an external benefit (such as an increase in security or capability).

In the defence domain these trade-offs are typically made via consideration of a wider system of systems, with the argument being that exposing soldiers to operational risk reduces the risk faced by civilians should the operational task not take place. Similar arguments are made for the risk posed by simulation training systems [12]. These training systems pose a certain degree of risk to soldiers, but this is offset by a decrease in risk when the wider operational environment is considered (trained soldiers are assumed to face less operational risk). In [12] [17], we propose that a similar approach could be used to trade one risk off against another. This would allow – for example – an increase in risk to pedestrians from the introduction of AVs to be justified by an equivalent or greater decrease in risk to other drivers. Such a trade-off would need to be explicitly integrated into the safety case, and made clear to all stakeholders.

A confounding factor is that, because ALARP does not recognise risk-benefit trade-offs, safety professionals are not accustomed to, or experienced in, making these. This problem is exacerbated when the difficulty of communication with stakeholders is considered. As we stated in Section 4, for consent to risk to be meaningful, all particulars of the risk and risk exposure must have been communicated. We would recommend that risk trade-offs be explicitly included within the safety case for AVs, and potentially considered as part of the certification.

5 Next steps

Public perception must also be considered in terms of next steps, and what AV manufacturers may need to do prior to introduction and acceptance of AVs. Public road testing is sometimes thought of as evidence of safety. However, this is not the case: such systems must be safe prior to deployment on the road. This means that V&V of the system (including

testing and demonstration) must be carried out within a controlled environment. The problem of ensuring that this environment adequately replicates a public road then becomes one of primary concern.

Another crucial issue is one of explicit communication with stakeholders, and with identifying the assumptions around ethics which have been made. A safety case which includes these assumptions would go some way towards increasing transparency. In [17] we recommend the explicit inclusion of an "ethics case": a supplementary case which considers the ethics embedded within the functionality of an AV.

Acknowledgements

The material for this white paper has its origins in the research performed and the workshops organised by the Transport Systems Catapult on behalf of the Department for Transport.
The authors wish to thank the SCSC Safety of Autonomous Systems Working Group.

References

[1] Harris, C., Pritchard, M., Rabins, M. "Engineering Ethics: Concepts and Cases", *Wadsworth, Cengage Learning*, 2009.
[2] MIT, "MIT Moral Machine", http://moralmachine.mit.edu/, 2017.
[3] Goodall, N. "Machine Ethics and Automated Vehicles", *Road Vehicle Automation*, pp 93 – 102, 2014.
[4] Health and Safety Executive, "Reducing Risks, Protecting People", 2001.
[5] Goodall, N. "Ethical Decision Making During Automated Vehicle Crashes", *Transportation Research Record Journal of the Transportation Research Board*, 2014.
[6] Lin, P. "Why Ethics Matter for Autonomous Cars", *Autonomous Driving*, pp 69 – 85, 2015.
[7] Gerdes, J., Thorntson, S. "Implementable Ethics for Autonomous Vehicles", *Autonomous Driving*, pp 87 – 102, 2016.
[8] International Standard for Organisation, "Road Vehicles – Functional Safety", *ISO 26262*, 2011.
[9] IEEE Global Initiative. "Ethically Aligned Design", *IEEE Standards v1.0*, 2016.
[10] Health and Safety Executive, "Safety Assessment Principles for Nuclear Facilities", 2006.
[11] Office for Nuclear Regulation, "Guidance on the Demonstration of ALARP", 2013.
[12] Menon, C., Bloomfield, R., Clements, T. "Interpreting ALARP", *Proceedings of the 8th IET International System Safety Conference*, 2013.
[13] Johnson, C. "Barriers to the Use of Intrusion Detection Systems in Safety-Critical Applications", *SAFECOMP*, 2014.
[14] Royal Academy of Engineering, "Statement of Ethical Principles", 2017.
[15] IET, "Rules of Conduct 2015", 2015.
[16] Habli, I., Kelly, T., Macnish, K., Megone, C., Nicholson, M., Rae, A. "The Ethics of Acceptable Safety", *Proceedings of the 23rd Safety-critical Systems Symposium*, 2015.
[17] Menon, C., Alexander, R. "A Safety-Case Approach to Ethical Considerations for Autonomous Vehicles", The Ethics of Acceptable Safety", *Proceedings of the 12th International Conference on System Safety and Cyber Security*, in press.
[18] Gartner. "Hype Cycle Methodologies", http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp, 2017.