A safety-case approach to the ethics of autonomous vehicles

Catherine Menon[a]* and Rob Alexander[b]

[a]*School of Computer Science, University of Hertfordshire, Hatfield, UK;* [b]*Department of Computer Science, University of York, York, UK*

a* c.menon@herts.ac.uk

b rob.alexander@york.ac.uk

Catherine Menon is a senior lecturer in the Adaptive Systems group at the University of Hertfordshire.

Rob Alexander is a lecturer and researcher in high-integrity systems engineering at the University of York.

A safety-case approach to the ethics of autonomous vehicles

Autonomous vehicles (AVs) have significant ethical and safety implications. Questions of informed consent and risk acceptance are of primary importance, as is an explicit identification of the ethical principles underlying these decisions. In this paper we present a method for translating ethical imperatives into design decisions and safety management practices. We demonstrate the use of an assurance case structure to justify the integrity of this translation. The assurance case combines mathematical and technical evidence with a compelling and comprehensible argument structure, and provides a means to demonstrate that design and safety management decisions reflect an identified ethical position.

Keywords: safety; ethics; autonomous vehicles; risk

1. Introduction

Autonomous vehicles (AVs) are increasingly being presented as the future of transport on public roads. The worth of the Connected and Autonomous Vehicle (CAV) market in the UK by 2035 is estimated to be 28bn (Transport Systems Catapult [TSC], 2017), and typical estimates for the first occurrence date of fully-autonomous vehicles on UK roads are in the mid-2020s (UK Government Department for Transport [DfT], 2015). While such forecasts are undoubtedly optimistic, they illuminate the continuing trend towards increased autonomy in this domain (Anderson & Anderson, 2007). Real-world trials of AVs are also underway in several countries, including the UK ((UK Government Centre for Connected and Autonomous Vehicles [CCAV], 2018), (UK Autodrive, 2017), (Venturer, 2016), (Venturer, 2017)), the US ((Waymo, 2018), (Uber, 2018)), Singapore, Japan and Europe (CCAV, 2018).

Although the technological capability to develop AV systems is developing quickly, ethical considerations remain an important societal barrier to their acceptance (UK Autodrive, 2017). This is exemplified by the trolley problem (Foot, 1967), which

in its original form presents an ethical dilemma in which a runaway train is on course to kill five people. A bystander is given the choice to let the train continue on its course or to divert it onto a different track where only one person will be killed. Although the trolley problem is a hypothetical question only, there is mainstream public interest in how AVs should behave in similar situations where harm to at least one person is inevitable (MIT, 2018).

It is arguable that the trolley problem has dominated the ethical debate around AV introduction to the exclusion of other, more nuanced, ethical considerations (Goodall, 2016). These other considerations include the balance and types of risk considered acceptable for AVs to present, the distribution of these risks across different sections of society, concerns over developer culpability and liability (Anderson, Nidhi, Stanley, Sorenson & Oluwatola, 2014), the environmental implications of AV introduction (Fagnant & Kockelman, 2014) and the impact on transport efficiency.

In this paper we build on existing work in risk distribution ((Menon & Alexander, 2017), (Menon, Bloomfield & Clements, 2013), (Menon & Alexander, 2018)) to present a methodology for translating from nuanced ethical considerations into safety requirements. This translation embeds explicit ethical principles within a *structured assurance case,* and uses the framework of this assurance case to demonstrate that the safety requirements corresponding to these ethical principles are satisfied by the AV. Structured assurance cases are well established in the area of safety argumentation ((Bloomfield & Bishop, 2010), (Kelly & Weaver, 2004), (Hawkins, Habli, Kelly & McDermid, 2013)) where they are used to provide a rigorous justification of the safety of a system. Our proposed approach uses these assurance cases within the ethical domain, using their structure to ensure that ethical principles are described, justified and implemented within the AV design.

Our approach does not attempt to prescribe a set of recommended ethical principles for AVs, but rather focuses on the translation of ethical principles into safety requirements. As such, the framework we present is broadly applicable across a range of credible ethical motivations, some of which are presented as examples in Section 5. However, our framework does assume compliance with the UK Human Rights Act (UK Government, 1998) and the other national implementations of the European Convention on Human Rights. In addition, we restrict our discussion throughout this paper on AVs which perform the whole of the driving task. These correspond to SAE Level 4 or Level 5, using the levels of automation identified and discussed in (SAE, 2018).

In Section 2 we present the ethical and safety background relevant to AVs, highlighting where existing work does not fully facilitate transformation of ethical concerns into safety requirements. Section 3 describes how these ethical concerns impact our judgements around risk acceptance and risk balancing. Section 4 provides a taxonomy of acceptable risk profiles which each represent a different ethical perspective on risk acceptance. Section 5 shows how the risk profiles can be used to translate ethical imperatives into specific safety requirements, using assurance case patterns to argue that these are satisfied. Section 6 provides discussion and Section 7 concludes.

2. Background

2.1 Structured assurance cases

Within safety engineering, *structured assurance cases* are used to present a compelling, credible argument that a system is safe in a given context ((Bloomfield & Bishop, 2010), (Kelly & Weaver, 2004), (Kelly, 2007), (Object Management Group [OMG], 2019)). An assurance case consists of a set of claims about the system, such as a claim that all hazards have been identified, or that the failure rate of the system is below a

certain threshold. These claims are supported with evidence, and with an argument that the evidence is sufficient to provide confidence in that claim. Claims can be broken down into sub-claims, and typically several pieces of evidence are needed to provide confidence that a claim has been satisfied.

Assurance cases allow safety management decisions to be interrogated (e.g. by a regulator) and defended. The adequacy of the argument is critical, and assurance cases typically attempt to satisfy certain principles within their overall argument construction. Arguments built on these principles all have certain aspects in common, and as such the principles provide a *template pattern*, or a recommended method of constructing the argument in order to minimise the chances of introducing a logical, technical or semantic error. For example, safety standards within the defence domain (UK Ministry of Defence [MOD], 2017) require that arguments in assurance cases demonstrate satisfaction of the following four principles:

1. Safety requirements are defined to address the system's contribution to hazards
2. The intent of the safety requirements shall be maintained throughout requirements decomposition
3. The safety requirements shall be satisfied
4. Hazardous behaviour of the system shall be identified and mitigated

When constructing the argument, developers can define claims which correspond to each of these principles. (In practice, a number of claims might be required in order to support each principle, and each claim might also support multiple principles.) A significant body of work already exists around the construction and discussion of template argument patterns, with common questions being focused on how to create template patterns which eliminate logical fallacies, unjustified assumptions, weakened conclusions or other similar faults ((Bloomfield & Bishop,

2010) (Kelly, 2007), (Hawkins, Habli, Kolovos, Paige & Kelly, 2015), (Common Criteria, 2007)).

2.2 Ethics and AVs

While there is currently no regulatory or legislative barrier to the testing and deployment of AVs in the UK (DfT, 2015), there are significant ethical and safety challenges. AVs have the potential to cause harm in all the ways that traditional cars do (e.g. collisions, harmful emissions, impacts on road efficiency), as well as via novel pathways. For example, cyber-security is a significant issue (DfT, 2017), (UK Autodrive, 2018) as AVs are vulnerable to being controlled by malicious third parties in a way that traditional vehicles are not.

More generally, a number of high-profile accidents (albeit at lower SAE levels) have illustrated some of the safety and technological challenges related to automation of the driving task. Misidentification of a pedestrian coupled with the removal of emergency braking resulted in a fatal accident for Uber in Tempe in 2018 (National Transportation Safety Board [NTSB], 2018), while a user relying on the automation beyond its stated capabilities led to a fatal Tesla accident in 2016 (NTSB, 2017). Factors in other fatal Tesla accidents have been identified as misidentification of road features and road blockages by the AV (NTSB, 2018b), (Boudette, 2016).

This potential for fatal accidents illuminates the ethical challenges connected with the operation of AVs. These accidents affect the general public's perception of AV risk as well as giving rise to discussions around the adequacy of mechanisms in place to reduce this risk. Such discussions must consider the selection of risk criteria, the acceptability of residual risk associated with the AV and the extent to which users have consented to bear this risk. Existing work on ethical design standards, such as (IEEE Global, 2018), examines high-level ethical concerns by looking at the societal benefits

and concerns around autonomous systems in general. Other work, such as (UK Autodrive, 2017b), focuses on the different ethical factors with the potential to influence the eventual behaviour of an AV.

Some of these ethical factors are the "human values", such as the AV developers' desire for fairness and the AV passengers' desire for personal autonomy (Thornton, 2018). The first of these could lead to developers preferring algorithms which prioritise polite and non-aggressive behaviour of the AV, while the second could lead to implementation of customised AV behaviours which allow passengers to choose the preferred style of driving (Kuderer, Gulati & Burgard, 2015). This has ethical implications in itself, in that the developers may choose to limit the choices of driving style to those which are non-aggressive, thereby depriving the passenger of some personal autonomy.

In addition to these "human values", there are less altruistic factors which may also influence the eventual behaviour of an AV. Self-interest and commercial competitiveness are likely to be relevant, given the results presented in (UK Autodrive, 2017b), which identify that potential customers prefer the AV to prioritise the safety of its passengers over any third parties. More generally, self-interest could lead to developers making choices about the AV behaviour specifically to reduce their culpability in the case of an accident (Shalev-Schwartz, Shammah & Shashua, 2017)

In terms of the behaviours which result from ethical choices, (Gips, 1995) and (Wallach & Allen, 2008) explore how the design of an autonomous system is affected by the extent to which ethical reasoning and capacity is embedded within it. (Dennis, Fisher, Slavkoviv & Webster, 2004) presents formal verification that a high-level ethical policy is satisfied by the eventual behaviour of the system. A general architecture for a robot capable of modelling its own actions using simulation and

predicting the ethical consequences of these is discussed in (Winfield, Blum & Liu, 2014) and (Vanderelst & Winfield, 2018) while (Arkin, Ulam & Wagner, 2012) considers simulation of a robot with an ethical framework.

2.2.1 Revisiting the trolley problem

Much existing public discussion of AV ethics focuses on the trolley problem, positing a situation in which an AV must choose which of two pedestrians to collide with. The trolley problem is often cited in public media as an illustration of AV safety issues, and public debate is typically focused around this problem or equivalent variants (MIT, 2018). However, real-world instances of the trolley problem are rare, and as typically discussed it assumes a level of engineering capability that is infeasible (UK Autodrive, 2017b), (Goodall, 2016).

A more realistic variant of the trolley problem considers how the AV can act in any given situation to minimise the overall risk (UK Autodrive, 2017b). This requires the AV to accurately estimate its own operational capacity and to adjust its behaviour accordingly (Nilsson, 2018). For example, in (Lin, 2015) the example is given of an AV driving closer (within its lane) to a smaller car on its left than to a truck on its right. This choice reduces the risk to the AV, as a collision with a small car is safer for the AV occupants than a collision with a truck. Another AV chooses differently, driving closer to the heavier vehicle with more effective safety systems (Lin, 2015). This second AV is optimising its driving position to reduce the overall risk it poses to other road users, as if it collides with the truck this is less likely to result in injuries than a collision with the car would be. A further variant on this situation is discussed in (Gerdes & Thornton, 2016). Other proposed situations include an AV choosing a “sacrificial” path, such as placing itself to block the trajectory of a runaway vehicle (Lin, 2015)

3. Translating ethics into design

Considered as a body, these works presented in Section 2.2 provide two important foundational results: how to work towards generating a set of ethical principles and how to identify the AV behaviours resulting from these. However, they do not provide a general mechanism for specifying the ethical principles, documenting the translation of these into AV behaviours, or justifying that all the identified ethically-motivated behaviours are in fact performed.

Formal verification, as shown in (Dennis et. al., 2004), (Winfield et. al., 2014), (Vanderelst & Winfield, 2018), is a valuable contribution in this area. However, formal verification of the entirety of a complex safety critical system – such as an AV – has historically been considered infeasible due to cost, technical limitations, and perceived difficulty ((Liu, Stavridou, & Duerte, 1995), (Knight, 2002), (Yoo, Jee & Cha, 2009)). Moreover, such verification is also intended to demonstrate correctness according to the specification only, and therefore does not consider wider issues of overall risk reduction, regulatory compliance or errors due to requirements elicitation and environmental change. Consideration of these issues is a legal requirement for safety-critical systems (Health and Safety Executive [HSE], 2001).

Consequently, there is a need for a framework which specifically examines the translation of ethical principles into AV behaviours, while considering the wider principles of risk management of these behaviours. Our framework addresses this by using structured assurance cases to fully express the ethical and risk requirements of the system, and show how these requirements have been derived and met.

We begin by defining some terminology for use throughout. In the more nuanced trolley problems discussed in Section 2.2.1, the behaviour of the AV is motivated by the intent to reduce risk, whether this be to its own occupants or to other

drivers on the road. Such a motivation, of course, is more properly ascribed to the AV developers rather than to the AV itself. This distinction highlights two interrelated areas of application when considering the ethics of AVs: *implemented ethics* (the ethics embedded within an autonomous system and realised in its behaviour) and *engineering ethics* (the ethical principles and codes of practice followed by engineers). Making this distinction allows us to interrogate the ethical behaviour of the AV without necessarily considering the professional conduct of the developers, and vice versa.

3.1 Engineering ethics and implemented ethics

Engineering ethics refers to the professional ethical principles which are followed by the developers of the AV during development work. These principles may be represented by professional codes of conduct (Royal Academy of Engineering [RAEng], 2017) as well as more general informal undertakings (Martin & Schinzinger, 2005).

Such ethical principles typically include criteria such as honesty, integrity, respect for law and the public interest, accuracy, rigour, fairness and objectivity (RAEng, 2017). However, they do not in themselves constrain the behaviour of any resulting system on ethical lines. It is, however, plausible that following a code of engineering ethics should prevent the developers from knowingly designing a system that contravenes established ethical foundations (e.g. the Human Rights Act (UK Government, 1998)).

Implemented ethics refers to the ethics embedded in the behaviour of the AV itself, sometimes referred to as the "moral algorithm" (UK Autodrive, 2017b) or the "machine ethics" of the AV. The implemented ethics determine how multiple risks are balanced against each other, and how safety risk is balanced against considerations of security, privacy, trust and capability. Different societies and stakeholders will differ in

their criteria for what behaviour is considered ethically acceptable, and this will also vary across environments and domains of use.

3.2 AVs and ethical risk reduction

Arguments have been put forward (Kalra & Groves, 2017). that AVs should be introduced as soon as their safety record is slightly better than traditional vehicles. Such studies estimate 500,000 fewer overall road fatalities over a fifty-year time frame compared to a conservative policy of AV introduction.

However, looking only at overall road fatalities obscures the distribution of such fatalities, which is crucial from both an ethical and safety perspective. The introduction of AVs may transfer risks even while reducing overall risk. That is, AVs may change how different classes of people (e.g. human drivers, passengers, pedestrians) are differentially exposed to risks. Any risk transfer also raises the question of risk consent, and whether all affected parties have agreed to the redistribution of risk.

A complicating factor here is that informed consent requires an accurate perception of the risk posed by the AV. Generally, perception of risk posed by a machine (AV) vs that posed by a human (driver) is not uniform across society (Kim & McGill, 2011), and in particular, those likely to perceive that AVs pose a lower risk than human drivers are, paradoxically, those most likely to bear a greater proportion of this risk due to their comparative lack of economic power.

Another complicating factor is the potential for AVs to contribute to risk indirectly, as participants in the wider road network. For example, if AVs cause more traffic jams, they may delay emergency vehicles, which may indirectly cause harm. As above, this indirect harm may not be equally distributed across the population and, moreover, may not be equally perceived by different segments of society.

A third factor in considering whether the risk posed by AVs is acceptable is the timing of its behavioural decisions. Unlike a traditional car — where the decisions during a crash are made in a time-critical frame — the intelligence in AV decisions is sited during design and implementation. Removing these decisions from a time-critical period argues that the resultant actions taken should be measurably better: that is, that they should reduce the risk posed by an AV when compared to a human driver (Groves & Kalra, 2017). This implies that it may not be acceptable for an AV to be merely "as good as" a human driver ((Kalra & Paddock, 2016), (Holloway, Knight & McDermid, 2014)), but that in order for AVs to achieve societal acceptance they must present a significantly lower risk than human drivers do.

The introduction of AVs which perform the majority of the driving task also transfers ownership of much of the risk associated with this task. Currently the driver owns much of this risk, system failure notwithstanding. However, for AVs at SAE levels 4 and 5, it is possible that the developer will own nearly the entire risk associated with the minute-by-minute driving decisions within the operational design domain. Although the human passenger may still provide input into route choices, customisation of driving techniques and overall usage, this would mean that AVs would be associated with a significant ethical responsibility borne by an individual or entity – the manufacturer and developer – not personally exposed to the resultant risk.

4 Ethics and risk balancing

As discussed in Section 3.2, one of the fundamental issues around ethics and safety of AVs is the question of risk transfer, or risk balancing. It is the redistribution of risks consequent on introduction of AVs which throws up the most complex ethical challenges.

There is a legal requirement in the UK for the overall risk associated with a system to be reduced As Low As Reasonably Practicable (ALARP). The Health and Safety Executive provides guidance (HSE, 2001) for good practice in reducing risk ALARP and for demonstrating this.

For a minority of systems, the system risk can be reduced ALARP by reducing the risk from each individual hazard ALARP. However, in cases where risk has been transferred and redistributed (as with AVs), the situation is more complex. It is possible to reduce a single risk at the cost of introducing another, or introduce a risk mitigation which affects multiple risks at once. This means that it is possible for multiple different system designs to all be ALARP, but for each to provide a different balance amongst the individual system risks (Menon et. al., 2013). This can occur under the following circumstances:

- When developers have not identified a complete list of hazards. This is relatively likely during the initial deployment of AVs, as safety engineers will not have all their usual tools for identifying and understanding hazards due to the lack of established good practice and historical data.
- When there are interdependencies between hazards which are not adequately accounted for. In these situations, some risks may be accounted for twice, giving a false idea of the overall risk associated with the system. There may also be interdependencies between systems and common mode failures which increase the complexity of combining risks. Given the relative novelty of AV technology and risk management, this is likely to be a significant issue.
- When multiple risk mitigations are not independent, and a mitigation for one risk potentially increases other risks. For example, increasing the sensitivity of algorithms which detect an object mitigates against the risk of failure to detect,

but increases the risk of erratic driving (the AV may brake unnecessarily to avoid a “phantom” object) and therefore the likelihood of a collision.

- When a single risk mitigation affects multiple hazards. In this case, the cost of the mitigation can be amortized over all the hazards and the resulting cost judged reasonably practicable, where it would not when assessed against each hazard individually.
- There are limited resources subject to a threshold effect of aggregation. For example, in a SAE level 3 vehicle (i.e. one with supervising driver), operator attention may be a mitigation against several hazards. When these hazards present themselves simultaneously, the operator may be overwhelmed.

In circumstances where multiple different designs all present an overall ALARP system risk, selecting any one of these designs represents a risk distribution choice. That is, each design provides a different balance amongst the individual system risks, “trading off” an increase in one risk for a decrease in another. This is an established practice in the nuclear domain, with standards such as (HSE, 2006) and (Office for Nuclear Regulation [ONR], 2018) emphasising the need to balance individual risks within a system. However, outside the nuclear domain many safety guidance documents provide little information on risk balancing and risk transfer, or require an explicit justification of any risk trade-offs.

4.1 Ethical motivation for risk balancing

Where multiple different system designs all offer an overall ALARP risk, the final choice may be made affected by a number of factors including cost, technical capability, resource availability and ethical imperatives. The cheapest design may be chosen, or the design which can be most easily implemented with the resources and technology at

hand. In these cases the justification is relatively simple and has long been part of standard project management techniques (Office of Government Commerce, 2019).

However, the ethical principles behind selecting a design are not generally explicitly discussed. Historically, there has been no mechanism to record ethical imperatives which result in one risk being considered to be more important than another. AVs in particular are vulnerable to the impact of such "hidden" ethical priorities, as exemplified in the trolley problem. In its most simple form, where the AV must choose between hitting one person or another, the developer might use ethical principles to decide which of these is preferable. These will result in different system designs, both of which present the same overall system risk.

The ethical complexities around AV operation and use have been discussed in Section 2. Because these have the potential to result in different distributions of risk, we consider that there is an obligation to provide information to the general public about any ethically-motivated risk trade-offs that have been made. Visibility of this information is necessary if affected stakeholders are to provide informed consent to the risks that they bear as a result. For example, where a developer has chosen a system design for the AV which prioritises the safety of passengers over pedestrians, this decision – motivated by the ethics of self-interest – must be made clear to both pedestrians and passengers. The framework we introduce in the following sections will ensure that such risk trade-off decisions are documented and justified, and any implications for the design of the AV are understood.

4.2 Refinement of risk factors

In order to provide explicit identification of those risk trade-offs which have been made in response to ethical imperatives, we first consider the breakdown of risk.

Risk is typically calculated to be dependent on both the outcome and the likelihood of this outcome

Risk = harm * likelihood

However, when assessing whether a risk trade-off is acceptable, we need greater transparency into this calculation. In particular, we need sufficient information to fully characterise and differentiate between risks, and therefore to justify why any given risk trade-off or balancing has been performed. With this in mind, we propose that assessment of risk for AVs should also consider the following factors.

a) The exposed population

We propose identifying several broad categories for the population who might be harmed by the AV. These include both direct stakeholders (e.g. someone the AV might collide with) and indirect stakeholders (e.g. someone affected by the emissions from an AV driving in a particular style). Categories of stakeholders should be chosen carefully, and some examples are likely to include:

- The AV passenger
- Other road users (e.g. other drivers, motorcyclists etc.)
- Pedestrians and cyclists
- Those living near the road

b) Area of impact

This defines the area in which the harm is experienced. The area might refer to the geographic area (the harm manifests close by vs harm manifesting further away) or the time taken for harm (the harm manifests straight away vs harm manifesting several years later). Some examples are:

- A crash causes geographically local harm (it harms pedestrians only in the immediate area)
- A road bottleneck causes geographically wider harm (the traffic jam delays an ambulance several streets away, causing harm to the patient)
- A crash causes harm that is close in time (the harmful effects of a crash are typically experienced immediately)
- Cancer from use of carcinogenic AV fuel is distant in time (cancer may take years to develop)

c) Causes

There may be specific risk causes which are of interest to an AV manufacturer. For example, risks due to cybersecurity failings may be of interest to a manufacturer developing AVs which rely on connected communication. Similarly, some risk causes result in greater reputational damage than others, including risks caused by software components which have been implicated in previous crashes.

4.3 Risk profiles and ethical positions

In this section we extend (Menon et. al., 2013) to present a taxonomy of ten different ethically-driven approaches to risk reduction choices. Each of these approaches describes a potential way to balance individual risks to achieve an ALARP system risk. These approaches provide a means of interpreting ethical imperatives from the perspective of safety management, and of making the risk trade-offs and balancing explicit. We will refer to these approaches as risk profiles.

Eight of the profiles correspond to the factors presented in Section 4.2 as constituting ethically-relevant characteristics of AV risk: likelihood, severity, exposed population, area of impact and causes. In addition to this, we consider two risk profiles

which correspond to techniques for balancing risks. This set of profiles is not intended to be exhaustive, but rather to serve as exemplars for the types of risk reduction decision which may be made, and the motivations behind these.

Risk profile name	***Description***
Likelihood risk profile	This prioritises the reduction of risks considered most likely to occur.
Severity risk profile	This prioritises the reduction of catastrophic events, which may not be very likely but would result in numerous fatalities (e.g. multi-car collisions).
Exposure risk profile	This prioritises the reduction of risks which are likely to result in harm to a certain defined segment of the population (e.g. children, pedestrians).
Local risk profile	This prioritises the reduction of those risks which are likely to result in geographically local harm over those risks which are more likely to result in harm at a greater distance, or where the harmful effect is spread over a wider aspect of the road network.
Global risk profile	This prioritises the reduction of those risks where the harm caused eventuates at a geographical distance, and most likely within a wider System of Systems (SoS).
Time-critical risk profile	This prioritises the reduction of risks which may lead to immediate harm over the reduction of risks where the harm occurs at some point in the future.
Long-term benefit risk profile	This prioritises the reduction of the long-term overall risk over the reduction of short-term, immediate risks.
Risk causes profile	This prioritises the reduction of risks that are due to particular identified causes, such as security flaws or malfunctions in a particular component.
Fairness in improvement risk profile	This prioritises achieving an equal decrease in risk for all individual risks.
Fairness in outcome risk profile	This prioritises achieving risk reduction that results in a similar level of risk for all individual risks.

Table 1: Risk profiles and descriptions

4.3.1 Risk profile selection guidance

Because of the complexity of the ethical challenges around AVs, it may be the case that multiple risk reduction profiles must be combined in order to capture all ethical imperatives. The following table identifies some typical scenarios and the appropriate risk reduction profile(s) for each. For each, we provide a justification as to why these

are the appropriate risk reduction profile(s) and a description of some possible implementation actions the developers might take in line with these profiles. We note that these actions are not the only way a developer might implement this risk reduction profile, and emphasise that these are to be considered as illustrative examples only.

The scenarios have been selected by conducting a survey of existing work to identify common ethical dilemmas and legal concerns ((Anderson & Anderson, 2007), (Fagnant & Kockelman, 2014), (Arkin et. al., 2012), (Gerdes & Thornton, 2016), (Kalra & Groves, 2017), (Kalra & Paddock, 2016), (Arman, 2018), (Thornton, 2018)). The selection and combination of the correct risk profiles for a given situation remains a complex question, however, and will be the subject of future validation work.

Scenario	***Risk reduction profile***	***Justification and implementation***
For reasons of self-interest (reputational damage) developers want to reduce the number of accidents the AV is involved in.	Likelihood	**Justification:** A likelihood risk profile prioritises the reduction of less harmful but more common risks **Implementation:** The drivers incorporate a "steady driving" style which reduces erratic driving at the potential cost of including some rare false negative results in object identification
For reasons of self-interest (insurance) developers want to reduce the number of accidents involving the AV which result in one or more fatalities.	Severity	**Justification:** A severity risk profile prioritises the reduction of those collisions which result in the most severe consequences. **Implementation:** The developers achieve this by increasing the sensitivity of object detection, at the risk of decreasing the efficiency of driving (due to an increased number of false positives)
There is reputational concern around being the first AV to potentially injure a child.	Exposure	**Justification:** An exposure risk profile is appropriate in situations in which there is ethical or reputational concern about exposing a particular segment of the population to risk **Implementation:** The developers choose to prioritise the detection and avoidance of children over adults in the event of a trolley problem scenario

Scenario	*Risk reduction profile*	*Justification and implementation*
Out of altruism, the developers want to minimise disruption to those people living in an area where AVs can't yet travel, on the basis that these people receive no benefit from the AVs	Global	**Justification:** A global risk profile is appropriate when considering safety risks which are propagated throughout a system, and where the effect may be felt at a distance from the risk origin (e.g. traffic congestion, wider road safety). **Implementation:** The developers choose to implement travel algorithms which minimise disruption at the boundary of the AV-enabled areas which might propagate outside these.
Out of pragmatism, the developers want to minimise disruption within the AV-enabled areas only, on the basis that they have little information about road conditions outside these and little obligation to the population there.	Local	**Justification:** A local risk profile approach is most useful where the interaction of risks at the system of systems (SoS) level is either minimal or difficult to quantify. **Implementation:** The developers choose to implement travel algorithms which are best for the road network efficiency within the AV-enabled area, and do not consider the effect outside this
Out of concern for the environment, the developers want to meet emission and carbon-neutral targets.	Time-critical	**Justification:** A time-critical risk profile allows for the explicit balancing of risks where the harm eventuates immediately against risks where the harm eventuates in the future. **Implementation:** The developers may choose to use fuel components with a risk of lung-irritant emissions causing immediate harm rather than components with a risk of harmful environmental emissions (all else being equal)
The developers want to introduce AVs immediately for "the greater good", being willing to accept some immediate casualties if it results in a long-term reduction in fatalities	Long-term benefit	**Justification:** The long-term benefit approach is based on questions raised in standards such as (HSE, 2006), which allow for the possibility of accepting a higher short-term risk if this reduces the risk long-term. **Implementation:** The developers may push to introduce AVs before sufficient evidence exists to show that they are at least as safe as a human driver. This concept of accepting a short-term higher risk is not current good practice (Menon, 2017), and therefore should only be used where an appropriate justification can be made.

Scenario	*Risk reduction profile*	*Justification and implementation*
For reasons of self-interest (reputational damage), the developers want to avoid public concern that the AV can be "hacked" and controlled remotely.	Risk causes	**Justification**: A risk causes approach is appropriate when developers prefer to prioritise risks resulting from security violations over risks resulting from other concerns. **Implementation:** This may lead developers to choose a component which offers better security updates over one which offers better diagnostic accuracy (all else being equal).
The developers want AVs to offer a similar improvement in safety to every segment of the population, compared to human drivers, out of a sense of fairness.	Fairness in improvement	**Justification:** The aim of this approach is to reduce *all* risks associated with a system by a certain minimum amount, without considering the relative cost of these reductions **Implementation:** The developers may make choices that ensure that the risk posed by the AV to every party (cyclists, pedestrians etc.) is reduced by a broadly similar amount when compared to the risk posed by human drivers.
The developers want AVs to be significantly safer than human drivers for the most vulnerable road users: pedestrians, cyclists and motor-cyclists.	Fairness in outcome	**Justification:** The aim of this approach is to achieve a similar *level* of risk for all individual risks, by reducing the highest of these before addressing any lesser risks. **Implementation:** The developers may make choices that will ensure the AV prioritises detection and avoidance of these vulnerable road users over that of other AVs, in the event of a trolley problem scenario.
Developers want to reduce the number of economically disadvantaged people involved in road accidents.	Combined Exposure and Fairness in Outcome	**Justification**: These profiles together prioritise risk reduction to a certain segment of the exposed population who are most likely to be economically disadvantage (pedestrians, cyclists, motorcyclists – but *not* AV passengers), along with reduction of the highest risks (again, those faced by pedestrians, cyclists and motorcyclists). **Implementation**: The developers will prioritise detection and avoidance of pedestrians, accepting the increase in false positives and the subsequent erratic or inefficient driving.

Scenario	*Risk reduction profile*	*Justification and implementation*
Developers want to avoid giving passengers the choice to have input into any AV decisions, including routing and travel choices.	Combined Risk Causes and Fairness in Outcome	**Justification:** These profiles together prioritise reduction of risks from a specified cause (dynamic passenger input), along with reduction of the highest risks (resulting from the AV acting on input from uniquely unqualified sources such as child passengers). This is at the cost of marginally increasing the risk where the passenger is in fact competent to provide input and has observed some aspect of the environment (e.g. traffic jam) that the AV has not. **Implementation:** The developers may choose to provide only fully-autonomous modes of driving which do not permit any input from the passengers.
For reasons of self-interest (e.g. limit of liability and insurance), the developers want to avoid any accidents involving terrorist activity or more than one person	Combined Severity and Risk Causes	**Justification:** These approaches together prioritise reduction of risks from a specified cause (security violations) together with reduction of the highest severity risks (those leading to multiple fatalities) **Implementation:** The developers may choose to implement an autonomous mode that's operational only at low speeds (reducing the chance of a catastrophic crash), and also to push and install security patches as soon as these are released.

Table 2: guidance for selecting risk profiles

5. A structured assurance case for ethical principles

The risk profiles of Section 4 allow us to translate from ethical principles (e.g. "avoid harming children") to risk management decisions which affect the design and safety of the AV system ("choose components better able to detect and distinguish between children and adults"). In this way, the risk profiles encourage developers to make the impact of ethics on the design explicit, by requiring a description of the risk balancing and risk trade-offs which result from these ethics.

These risk profiles therefore provide us with a generalised mechanism to discuss the ethics embedded in the design, the impact of these ethics on design and safety decisions and the implications for risk management. Such a mechanism ensures transparency, and ensures that stakeholders have been supplied with sufficient information to consent to the proportion of risk that they bear, or to the imposition of risk that they will be placing on other people by using an AV. The risk profiles also allow traceability between ethical factors and design decisions, which is essential if sufficient confidence in the adequacy of the eventual AV behaviour is to be achieved.

However, the risk profiles alone provide only a textual description, without the ability to inform or link into the safety case. This can lead to information being lost or not updated as required, and adds complexity to assessment of AV safety. To address this issue, we present a methodology for linking the risk profiles and ethical information to the safety case by using a template pattern within a structured assurance case.

5.1. Principles of ethics assurance cases

As discussed in Section 2, it is common for structured assurance cases to be founded on the satisfaction of certain principles, and for these principles to be key to the safety argument construction. Within the defence domain, there are four key principles that must be satisfied within a safety argument (MOD, 2017). These have been introduced in Section 2, and we have chosen to extend these into the area of ethics to form the key principles of our structured assurance case dealing with ethics. These extended principles are as follows:

P1. Ethics requirements appropriate for AV development and operation shall be defined.

This requires that engineering ethics and implemented ethics

requirements should be explicitly defined, free from inconsistencies, and containing sufficient detail to allow the other principles to be met.

P2. The intent of the ethics requirements shall be maintained throughout decomposition.

This requires that the implemented ethics should be propagated throughout the design of the system and refined into lower-level requirements on design, implementation and risk management. The engineering ethics should be satisfied throughout the system lifecycle.

P3. Ethics requirements shall be satisfied.

This requires that the ethics requirements, both implemented and engineering, should be demonstrably satisfied and evidence provided to support this.

P4. The AV shall continue to be safe, and emergent behaviour of the AV which conflicts with the ethics requirements shall be identified and mitigated

This constrains emergent behaviour of the AV, either due to changes in the environment or to adaptive algorithms used within the AV software. Such emergent behaviours may not have been considered when specifying the original ethics and safety requirements, and this principle requires that evidence be provided to assure the continued safety of the AV even in a changing environment.

There is a further principle relating to confidence in (MOD, 2015), which has no immediate analogue to ethics, and which we do not develop further.

5.3. Ethics assurance case template

Using the principles P1 – P4 defined above, we can now construct a template pattern for constructing an argument which satisfies these. This template pattern will make explicit use of the risk profiles of Section 4 to translate between ethical requirements and safety / design requirements. The top-level claim is that the AV behaves – and continues to behave – in a manner which is ethically appropriate for its environment.

We present this template pattern in Figure 1 using a diagrammatic notation (Goal Structuring Notation (Assurance Case Working Group, 2018)), and discuss it textually below.

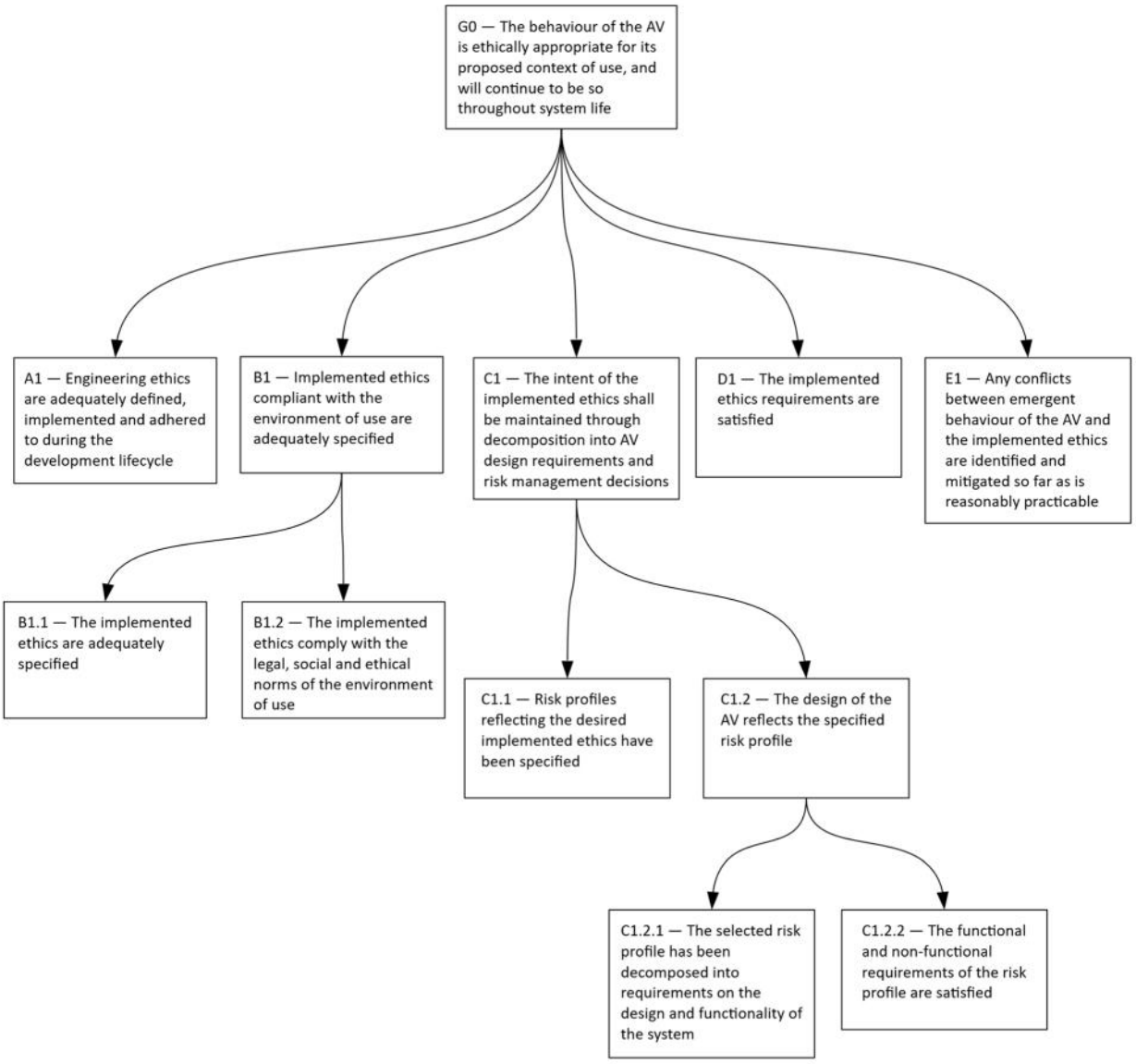

Figure 1: ethical assurance template argument pattern

Claim A1: *Engineering ethics are adequately defined, implemented and adhered to during the system lifecycle*

This claim partially supports P3 and requires AV developers to demonstrate compliance with an appropriate code of conduct, domain-specific good practice and existing ethical precedents. This claim might reasonably be supported with records from audits and artefacts from the development lifecycle, as well as documentation that developers have

followed processes defined in accordance with codes of professional conduct. This claim helps to provide confidence in the integrity of any lifecycle artefacts which are needed to support the top-level claim of Figure 1.

Claim B1: *Implemented ethics compliant with the environment of use are adequately specified.*

This claim supports P1, and is broken down into two sub-claims as follows:

Claim B1.1: *The implemented ethics are adequately specified*

Specification of the implemented ethics may be achieved via identification and citation of relevant items of regulation and policy, as well as the results of any public consultation or ethical objections already tabled. It may also be useful to specify aspects of implemented ethics via references to previous system designs, to academic papers, and to accepted good practice. The specification of the implemented ethics must be sufficient to address competing ethical motivations, and its adequacy must be explicitly justified.

Claim B1.2: *The implemented ethics comply with the legal, social and ethical norms of the environment of use*

The implemented ethics must be compatible with behaviour that would be reasonably expected by the general public for an AV operating within the stated environment. It should be noted that this does not necessarily imply an AV should behave in exactly the same way as a traditional driver (IEEE Global, 2018), but rather that the AV should act in a way that a traditional driver might plausibly expect from an AV.

Claim C1: *The intent of the implemented ethics shall be maintained throughout decomposition into AS design requirements and risk management decisions.*

This claim supports P2, and serves to translate ethical requirements into lower-level safety and risk management requirements which implement the ethical intent. It is broken down into two sub-claims as follows:

Claim C1.1: *Risk profiles reflecting the desired implemented ethics across different environments have been specified*

The risk profiles discussed in Section 4 provide a method of reflecting ethical perspectives in risk management, risk balancing and risk distribution decisions. Satisfaction of this claim requires that (a combination of) risk profiles have been defined for all relevant scenarios, environment and operational lifecycle phases.

Claim C1.2: *The design of the AV reflects the specified risk profile*

This claim is supported by an argument that the risk balancing inherent in the specified risk profile has been performed accordingly, via the translation of this risk profile into technical, safety and risk requirements. It is broken down into three further sub-claims as follows:

Claim C1.2.1: *The selected risk profile has been decomposed into requirements on the design and functionality of the system*

This claim is supported by referencing out to the design and implementation requirements specification, as well as to the safety case.

Claim C1.2.2: *The design and functional requirements of the system are satisfied.*

This claim is supported by referencing evidence provided within the safety case, which is the primary mechanism for demonstrating satisfaction of design and safety requirements.

Claim D1: *The implemented ethics requirements are satisfied.*

This claim partially supports P3 and requires identification of what the acceptable ethical behaviour of the AV might be. Natural-language interpretation of the ethics

requirements will help support this claim, as will a description of the functionality and behaviours which comply with these ethics. Supplementary supporting evidence will include system verification and validation of the derived requirements sourced from the ethical imperatives via risk profiles. Traceability between these derived requirements and any further lower-level requirements must be demonstrated, along with traceability between these derived requirements and verification artefacts.

Claim E1: *Any conflicts between emergent behaviour of the AV and the implemented ethics are identified and mitigated so far as is reasonably practicable.*

This claim supports P4 and requires an estimation of likely emergent behaviours and changes in the environment throughout the AV's lifetime. Support for this claim requires a gap analysis of potential environmental change, as well as of gap analysis between the behaviours which may potentially be learnt by AVs (via adaptive algorithms and continuous machine learning) and the behaviours which were "hard-coded" or scripted at deployment. Any conflicts between the ethics requirements and these new behaviours and environments must be identified and mitigated so far as is reasonably practicable.

6. Discussion

The template pattern in Section 5 allows developers to argue that the behaviour of the AV is ethically appropriate for its environment of use. It requires them to identify their underlying ethical imperatives and how they have translated these into design decisions by making use of the risk profiles. There is nothing in this template fragment itself to constrain the developers' ethical imperatives, so both "positive" (altruistic) and "negative" (self-interested) ethical imperatives can be represented. This is a deliberate decision to ensure that the structured assurance case has the capability to represent all

the relevant ethical aspects of the AV, rather than simply those which are the most palatable to stakeholders.

We do note, however, that under Claim B1, developers are required to justify that the implemented ethics comply with the legal, social and ethical norms of the environment of use. Where developers have prioritised self-interest or reputational damage (as shown in Section 4, e.g. “For reasons of self-interest, developers want to reduce the number of accidents the AV is involved in”), it is likely that the justification of the ethics behind this is comparatively weak. In such cases the subsequent AV behaviour (e.g. a “steady driving” style reducing common low-speed accidents) may be acceptable in itself, even if the ethics motivating it is not. By separating out the ethics and the behaviours – and producing a template argument pattern which requires developers to be explicit about both – we have ensured that stakeholder have transparency into both the design decisions and the underlying ethics which motivate these.

We have suggested possible items of evidence which might support each claim. These are intended as illustrative guidance only and should not be considered exhaustive. Template patterns do not typically constrain the evidence used in support of claims, leaving this to best practice and the engineering judgement of the developers. We do, however, note that – like safety – ethics is a limit concept (Kelly, Habli, Nicholson, Megone & Mcnish, 2014). It is therefore it is not possible to state definitively that an AV is “100% ethical”, and moreover, this will inevitably be a subjective judgment. When interpreting or instantiating the template pattern, the top-level claim should therefore be understood to apply only so far as is reasonably practicable. This parallels the As Low As Reasonably Practicable (ALARP) caveat for system safety claims (HSE, 2001).

7 Conclusions

In this paper, we have discussed how ethical factors can affect AV behaviour. These ethical factors range from altruism on the developers' part to self-interest, and can be complex and contradictory. They affect risk management decisions, including risk transfer, risk consent, risk acceptability and risk balancing. In particular, as seen in the trolley problem, they can transfer risk from one section of the population and impose it on another.

In order to make such ethical factors transparent, we have identified a methodology for explicitly translating these into safety and design requirements, using a structured assurance case. Such an assurance case enhances transparency, by ensuring that the underlying motives which have led to individual AV behaviours are identified, justified and explicitly discussed. In order to facilitate the translation of ethical imperatives into risk management decisions we have identified risk profiles. These describe approaches to risk reduction by considering those cases where risks are balanced against each other, and a small increase in one risk is accepted for a proportionate decrease in another.

We have presented a template argument pattern, which can be used within a structured assurance case to construct an argument that the behaviour of the AV is ethically appropriate. This template pattern draws on the risk profiles, requiring developers to examine and explicitly identify the underlying ethical factors which have motivated the AVs implemented behaviour. The template argument pattern allows for expression of both altruistic and self-interested principles, and links to the safety case to provide evidence that design decisions stemming from ethical requirements have been met.

We propose to expand this work in future to apply the assurance case structure to a working case study. This will allow us to develop further risk profiles in

conjunction with industry personnel. In order to achieve this, we will also consider formalising certain ethical principles, including the principle of double effect and Kantian ethics, along the lines of (Bentzen, 2015), (Linder & Bentzen, 2018). We will also seek to identify a representative set of consequentialist ethics relevant to AV introduction and operation, and demonstrate the feasibility of our assurance case structure to represent this.

References

Ahmad, A. (2018). Driverless cars: Autonomous driving, a case for greater good, *New York Times* report https://www.nst.com.my/cbt/2018/03/349120/driverless-cars-autonomous-driving-case-greater-good.

Anderson, J., Nidhi, K., Stanley, K., Sorenson, P., Samaras, C. & Oluwatola, O. (2014). *Autonomous Vehicle Technology: A Guide for Policymakers,* Rand Corporation.

Anderson, M., & Anderson, S. (2007). Machine ethics: Creating an ethical intelligent agent, *AI Magazine* 28, 15—26.

Arkin, R., Ulam, P. & Wagner, A. (2012)Moral decision-making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. In *Proceedings of the IEEE*, volume 100, 571--589.

Bentzen, M. (2016). The principle of double effect applied to ethical dilemmas of social robots. In *Proceedings of the 2nd International Conference on Robophilosophy*, 268—279.

Bloomfield, R. & Bishop, P. (2010). Safety and assurance cases: Past, present and possible future: an Adelard perspective. In Proceedings of the Eighteenth Safety-Critical Systems Symposium.

Boudette, N. (2016). Autopilot cited in death of Chinese tesla driver, *The New York Times*. https://www.nytimes.com/2016/09/15/business/fatal-tesla-crash-in-china-involved-autopilot-government-tv-says.html, 2016.

Common Criteria Management Board. (2007). Common Methodology for Information Technology Security Evaluation. CCMB-2007-09-004.

Dennis, L., Fisher, M., Slavkoviv, M. & Webster, M. (2004). Formal verification of ethical choices in autonomous systems, *Robotics and Autonomous Systems* 77, 1--14.

Donde, J. (2017). Self-driving cars will kill people. Who decides who dies? *Wired.* https://www.wired.com/story/self-driving-cars-will-kill-people-who-decides-who-dies.

Fagnant, D. & Kockelman, K. (2014). The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. In *Proceedings of the 93rd Annual Meeting of the Transportation Review Board*.

Foot, P. (1967). The problem of abortion and the doctrine of double effect, *Oxford Review* 5, 5—16.

Gerdes, J. & Thornton, S. (2016). Implementable ethics for autonomous vehicles, *Autonomous Driving* 87--102.

Gips, J. (1995). Towards the Ethical Robot. *Android Epistemology*, MIT Press, 243—252.

Goodall, N. (2016). Away from trolley problems and toward risk management, *Applied Artificial Intelligence*, vol. 30, no. 8, 810—821.

Groves, D., Kalra, N. (2017). *Autonomous Vehicle Safety Scenario Explorer*, Web-Only Tool, RAND Corporation, https://www.rand.org/pubs/tools/TL279.html.

Object Management Group. (2019) Structured Assurance Case Metamodel (SACM), Document Number 20190314. https://www.omg.org/spec/SACM/2.1/Beta1/

Hawkins, R., Habli, I., Kelly, T. & McDermid, J. (2013). Assurance Cases and Prescriptive Software Certification: A Comparative Study, *Safety Science*, 55--71.

Hawkins, R., Habli ,I., Kolovos, D., Paige, R. & Kelly, T. (2015). Weaving an Assurance Case from Design: A Model-based Approach. In *Proceedings of the 16th IEEE International Symposium on High Assurance Systems Engineering.*

Health and Safety Executive. (2001). *Reducing Risks, Protecting People*. http://www.hse.gov.uk/risk/theory/r2p2.pdf

Health and Safety Executive. (2002). T*he Precautionary Principle: Policy and Application.* http://www.hse.gov.uk/aboutus/meetings/committees/ilgra/pppa.htm

Holloway, C., Knight, J. & McDermid, J. (2014). Neither Pollyanna nor Chicken Little: Thoughts on the ethics of automation. In *Proceedings of the IEEE International Symposium on Ethics in Science, Technology and Engineering*, 1--7.

IEEE Global Initiative. (2018). *Ethically aligned design, v2.0.* https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

Kalra, N. & Groves, D. (2017). *The Enemy of Good - Estimating the Cost of Waiting for Nearly Perfect Automated Vehicles*, Technical Report RR-2150 RC, Rand Corporation. https://www.rand.org/pubs/research_reports/RR2150.html

Kalra, N. & Paddock, S. (2016). *Driving to Safety: How Many Miles of Driving Would It Take To Demonstrate AV Reliability?,* Technical Report RR1478, RAND Corporation, https://www.rand.org/pubs/researchreports/RR1478.html.

Kelly, T. & Weaver, R. (2004). The goal structuring notation — a safety argument notation. In *Proceedings of Dependable Systems and Networks 2004 Workshop on Assurance Cases.*

Kelly, T., Habli, I., Nicholson, M., Megone, C. & Mcnish, K. (2014). The ethics of acceptable safety. In *Proceedings of the 23rd Safety-critical Systems Symposium.*

Kelly, T. (2007). Reviewing Assurance Arguments — A Step-By-Step Approach. In *Proceedings of the 12th International Conference on Dependable Systems and Networks DSN* 84—95.

Kim, S. & McGill, A. (2011). Gaming With Mr Slot or Gaming the Slot Machine?, *Journal of Consumer Research*, Vol 38, No 1, 94--107.

Knight, J. (2002). Safety Critical Systems: Challenges and Directions. In *Proceedings of the 24th International Conference on Software Engineering*, 547--550.

Kuderer, M., Gulati, S. & Burgard, W. (2015). Learning Driving Styles for Autonomous Vehicles from Demonstration. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).*

Lin, P. (2015). Why Ethics Matters for Autonomous Cars. In *Autonomes Fahren*, Springer Vieweg, 69 –85.

Lindner, F. & Bentzen, M. (2018). A formalization of Kant's second formulation of the categorical imperative. In *Proceedings of the 14th International Conference on Deontic Logic and Normative Systems*, College Publications.

Liu, S., Stavridou, V., Duerte, B. (1995). The Practice of Formal Models in Safety Critical Systems. *Journal of Systems and Software*, 77—87.

Martin, M. & Schinzinger, R. (2005). *Ethics in engineering*, McGraw-Hill New York.

Menon, C. & Alexander, R. (2017). A safety-case approach to ethical considerations for autonomous vehicles. In *Proceedings of the12th International Conference on System Safety and Cyber Security.*

Menon, C., Bloomfield, R. & Clements, T. (2013). Interpreting ALARP. In *Proceedings of the 8th IET International System Safety Conference.*

Menon, C. & Alexander, R. (2018). Ethics and the safety of autonomous systems. In *Proceedings of the 26th Safety Critical Systems Symposium.*

Menon, C. (2017). *A White Paper Discussing Ethical Considerations for Autonomous Vehicles*, Technical Report, Transport Systems Catapult, https://ts.catapult.org.uk/intelligent-mobility/im-resources/research-papers/

MIT. (2018). *MIT Moral Machine*. http://moralmachine.mit.edu/.

National Transportation Safety Board. (2018). *Preliminary Report Highway HWY18MH010,* Technical Report, National Transportation Safety Board, https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY18MH010prelim.pdf.

National Transportation Safety Board. (2016). *Collision Between a Car Operating with Automated Vehicle Control Systems and a Tractor Semitrailer Truck Near Williston, Florida, May 7 2016,* Technical Report HAR1702, https://www.ntsb.gov/investigations/AccidentReports/Pages/HAR1702.aspx.

National Transportation Safety Board (2018). *Preliminary Report Highway HWY18FH011,* Technical Report HWYFH011, https://ntsb.gov/investigations/AccidentReports/Reports/HWY18FH011preliminary.pdf.

Nilsson, J. (2018). Safe self-driving cars: Challenges and some solutions. In *Proceedings of the 26th Safety Critical Systems Symposium*.

Office for Nuclear Regulation. (2006). *Safety Assessment Principles for Nuclear Facilities*. http://www.onr.org.uk/saps/saps2014.pdf

Office for Nuclear Regulation (2018). *Guidance on the Demonstration of ALARP*, http://www.onr.org.uk/operational/tech_asst_guides/ns-tast-gd-005.pdf

Office of Government Commerce. (2019). *Prince2, Project Methodology*, https://www.prince2.com/uk/prince2-methodology.

Royal Academy of Engineering (2017). *Statement of ethical principles.* https://www.engc.org.uk/media/2337/statement-of-ethical-principles-2014.pdf

SAE International. (2018). *J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.* https://www.sae.org/standards/content/j3016_201401/

Shalev-Schwartz, S., Shammah, S. & Shashua, A. (2017). *On a Formal Model of Safe and Scalable Self-driving Cars*, arXiv preprint arXiv:1708.06374.

The Assurance Case Working Group. (2018). *Goal Structuring Notation Community Standard (Version 2),* Technical Report SCSC-141B, The Safety Critical Systems Club. https://scsc.uk/scsc-141B

Thornton, S. (2018). *Autonomous vehicle motion planning with ethical considerations*, (Doctoral dissertation), Stanford University.

Transport Systems Catapult. (2017) *Market Forecast For Connected And Autonomous Vehicles.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/642813/15780_TSC_Market_Forecast_for_CAV_Report_FINAL.pdf

Uber Advanced Technologies Group. (2018). *Safety Report: A Principled Approach to Safety*. https://uber.app.box.com/v/UberATGSafetyReport

UK Autodrive. (2017). *Public Attitudes Survey.* http://www.ukautodrive.com/wpcontent/uploads/2017/08/Executive-Summary-FINAL.pdf.

UK Autodrive. (2017). *The Moral Algorithm.* http://www.ukautodrive.com/wp-content/uploads/2017/08/moral algorithm white paper A4 051216.pdf.

UK Autodrive. (2018). *CAV: A Hacker's Delight*, http://www.ukautodrive.com/wp-content/uploads/2018/01/Cyber security White paper A4 050917.pdf.

UK Government Centre for Connected and Autonomous Vehicles. (2018). *UK Connected And Autonomous Vehicle Research And Development Projects.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/737778/ccav-research-and-development-projects.pdf

UK Government Department for Transport. (2015). *The Pathway To Driverless Cars.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/446316/pathway-driverless-cars.pdf

UK Government Department for Transport. (2017). *The key principles of vehicle cyber security for connected and automated vehicles.* https://www.gov.uk/government/publications/principles-of-cyber-security-for-

connected-and-automated-vehicles/the-key-principles-of-vehicle-cyber-security-for-connected-and-automated-vehicles

UK Government (1998) *Human Rights Act.* https://www.legislation.gov.uk/ukpga/1998/42/contents

UK Ministry of Defence. (2017). *Defence Standard 00-56: Safety Management Requirements for Defence Systems*, Technical Report 00-56 Issue 7.

Venturer Cars. (2016). *Introducing Driverless Cars to UK Roads*. https://www.venturer-cars.com/wp-content/uploads/2016/08/VENTURER-Trial-1-Overview.pdf

Venturer Cars. (2017). *Interactions Between Autonomous Vehicles and Other Vehicles at Links and Junctions.* http://www.venturer-cars.com/wp-content/uploads/2017/11/VENTURER-Trial-2-Technical-Reportv2.pdf.

Vanderelst, D. & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition, *Cognitive Systems Research* 48 56--66.

Wallach, W. & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press.

Waymo. (2018). *Safety Report: On the Road to Fully Self-Driving.* https://storage.googleapis.com/sdc-prod/v1/safety-report/Safety%20Report%202018.pdf.

Winfield, A., Blum, C. & Liu, W. (2014). Towards an ethical robot: Internal models, consequences and ethical action selection. In *Advances in Autonomous Robotics Systems*, 85--96.

Yoo, J., Jee, E. & Cha, S. (2009). Formal Modelling and Verification of Safety-Critical Software, *IEEE Software*, 42—49.