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 process framework for producing an ethics assurance case, which can be used to translate ethical imperatives into design decisions and safety management practices. The process and resultant assurance case integrate ethical considerations into the wider engineering lifecycle, providing a tool 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). These forecasts confirm 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 (e.g. as differentiated by factors including age, economic position, gender) and the different appetites for risk over these same sections (Wang & Zhao, 2019), 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 process framework for producing an ethics assurance case. This assurance case can be used to translate nuanced ethical considerations into safety and design requirements, and to demonstrate that identified ethical principles – and the safety requirements stemming from these – are satisfied. 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 embeds an assurance case template within a process structure, thereby creating a

framework for explicitly describing ethical principles within the engineering lifecycle. The framework ensures that ethical principles are described, justified and implemented within the AV design.

Our framework does not 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 is applicable to a range of ethical, risk perception and decision-making theories, some of which are discussed in Section 2. The 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. 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 International, 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 describes how ethical principles may govern how risks are balanced, or traded off, against each other and introduces the concept of risk profiles to describe these trade-offs. Section 5 presents an ethics assurance case template pattern and a process framework for instantiating this template to produce an ethics assurance case, thereby enabling the translation of ethical principles into specific safety requirements. Section 6 concludes with a discussion and steps for further work.

2. Ethical and Safety Background Relevant to AVs

2.1 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.

This potential for AVs to cause harm means that there are significant ethical challenges connected with their operation. Public perception of AV safety must be considered, alongside the demonstrated safety of such systems and the adequacy of mechanisms in place to reduce the risk posed by AVs. Such discussions must also 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. We note that these are also ethical issues for the existing car fleet, in that drivers, pedestrians and other road users have not explicitly consented to the risk as currently posed by human drivers. It is arguable that a similar ethical argument should be made for the existing fleet (as an alternative to the introduction of AVs). Such an argument is beyond the scope of this paper, but we note it as a prevailing ethical challenge.

2.2 Risk Perception Theories

The study of risk perception and decision making is a wide field, drawing from cross-disciplinary domains including psychology, linguistics and engineering. In this section we present a number of risk perception theories which are particularly relevant

to the societal perception of AVs and their risk acceptability.

Expected utility theory (von Neumann, 1947) provides a model of risk perception which allows for an individual's risk appetite and risk aversion. These properties are crucial when considering the ethics of AV use, as an individual's perception of the novel risks presented by an AV will be affected by their risk appetite. As stated earlier, the current (human-driven) fleet also presents risks which have not necessarily been explicitly consented to, and the framing of the new AV risks as compared to the existing risks can have a significant effect on people's perception of these.

Both expected utility theory and prospect theory (Kahneman & Tversky, 1979) consider the framing of risks, but prospect theory allows for greater complexity around risk perception. In particular, it provides a method of addressing the Allais paradox (Allais, 1953), and allows us to model risk perceptions around specific AV scenarios more accurately. (Wang & Zhao, 2019) use prospect theory to discuss AV risk perception in more detail, noting that the demographics which have a higher appetite for risk (e.g. young, fully-employed, male, high-income) view new technology more favourably, while the risk-averse (poor, elderly) perceive AVs as presenting a risk which – when considered in the context of perceived benefits – they are not prepared to accept.

Regret theory (Bell, 1983) provides a model of risk perception and decision making which captures the effect of regret, compared to an alternative outcome. Regret theory allows for the consideration of regret-avoidance, together with risk aversion and risk appetite. (Somasundaram & Diecidue, 2015) use regret theory to consider risk attitudes in society, finding that feedback polarizes regret attitudes, and increases risk-seeking amongst those who are regret-averse. As before, an increase in risk-seeking can affect the perceived risks posed by AVs. It presents an ethical challenge, particularly as AVs pose a risk across

multiple segments of society, and there is the potential for an increase in risk-seeking by one segment to result in a consequently greater risk being posed to another.

Dual Process Theory (Baron, 1985; Evans, 1989) allows for the interaction between decisions made for fast, intuitive reasons and slower decisions based on reasoning. As with prospect theory, dual process theory emphasises the importance of attribute framing (Cokey, 2009; Reyna, 2004), which is particularly important for AVs given the complexity of the risks they present.

2.2.1 Risk perception of AVs

Existing work on ethical design standards, such as (IEEE Global, 2018) utilises risk perception theories to examine high-level ethical concerns around AVs, as well as providing an overview of 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. These ethical factors can be drawn from the ethical theories discussed above and include the "human values", such as the AV developers' desire for fairness and the AV passengers' desire for personal autonomy (Thornton, 2018). In practical, engineering terms 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). The concerns around how risks are framed is particularly relevant here, and several of the relevant ethical theories – including prospect theory and regret theory – identify framing as an important factor in willingness to accept a given risk.

In terms of the behaviours which result from ethical choices, (Gips, 1995; 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) present 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; Vanderelst & Winfield, 2018) while (Arkin, Ulam & Wagner, 2012) considers simulation of a robot with an ethical framework.

2.2.2 Revisiting the trolley problem

Much of the on-going public discussion of AV ethics focuses on the trolley problem, which posits a situation in which an AV must choose which of two pedestrians with which to collide. The trolley problem is often cited in public media as an illustration of AV safety issues, and public debate is typically focused around variants of this (MIT, 2018). However, real-world instances of the trolley problem are rare, and much of the public discussion around AVs and the trolley problem 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). Similarly, other work (Lin, 2015) considers the example 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 (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, the works presented in Section 2 provide two important foundational results: how to specify AV behaviours in individual scenarios, and how to assess AV behaviours resulting from given ethical principles. However, they do not provide a general mechanism for specifying the ethical principles, documenting the translation of these into AV behaviours, or justifying that ethically-motivated behaviours are demonstrated and will continue to be so.

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 inadequate requirements elicitation and environmental changes. Consideration of these issues is a legal requirement for safety-critical systems (Health and Safety Executive [HSE], 2001).

In the more nuanced trolley problems discussed in Section 2.2.2, 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 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. (As we discussed earlier, such explicit consent has not in fact been sought for the current fleet. People's decision to accept or reject the current fleet risk is largely intuitive, corresponding to many of the factors discussed in Section 2.2. The current method of formally adjudicating whether human intuition and judgement represent sufficient risk

mitigation given a particular accident scenario appeals to legal processes rather than ethics).

A complicating factor here is that informed consent requires an accurate perception of the risk posed by the AV. The decision-making theories of Section 2.2 identify a number of different motivating factors for decision-making under risk, but most compelling for AVs is that the perception of risk posed by the AV vs that posed by a human (driver) is not uniform across society (Kim & McGill, 2011; Wang & Zhao, 2019). A Singaporean study (Wang & Zhao, 2019) has shown that those who are more likely to benefit from AVs (high-income, urban-dwelling) are more willing to accept the risks, while a more general study of autonomous systems (Kim & McGill, 2011) shows that low-income segments of society are more likely to trust machines rather than interrogate the engineering and ethical principles behind these.

Another complicating factor is the potential for AVs to contribute to risk indirectly, as participants in the wider road network. For example, AVs which mis-identify emergency vehicles may not pull over for these vehicles, thereby indirectly causing harm. Similarly, AVs which learn to behave selfishly (e.g. failing to give way, aggressive driving) may also cause traffic congestion across the wider road network

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 possible outcomes of the resultant risks.

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 mitigating the risk from each individual hazard until it is 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 complete and valid 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; 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 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 those ethical imperatives which result in developers considering one risk to be more important than another. AVs in particular are vulnerable to the impact of such "hidden" ethical priorities of the designers. This is exemplified in the trolley problem, in its simplified form. The AV designer will use his / her ethical principles to decide whether the AV – guided by its programming – should impact one person or the other. Each of these choices will result in different system designs (because of the difference in behaviour), which present different risks to different people. However, the same overall system risk will remain the same whichever person is impacted.

Because the ethical complexities around AV operation and use 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 the ethically-motivated risk trade-offs that have made during development. Furthermore, where ethical considerations have led designers to consider one risk acceptable (while deeming

another equivalent risk not acceptable) and to have acted on this to inform the system design, this practice must be communicated to the general public. For example, where a developer has chosen a system design for the AV which prioritises the safety of passengers over pedestrians, this decision should be made explicit to both pedestrians and passengers. Visibility of this information is necessary if affected stakeholders are to provide informed consent to the risks that they bear as a result. Such visibility is, of course, lacking in the current fleet (pedestrians are unaware of the degree of self-interest of drivers), with current decisions around risk acceptance being made according to intuitive or personal ethical theories where in many cases insufficient information is provided for a full understanding.

4.2 Refinement of risk factors

When deciding which of two equivalent risks is preferable, or when performing risk trade-offs, the AV developer must consider more than the magnitude of the risks. There are other factors which may make one risk preferable (to the developer) than another risk of equivalent magnitude, in terms of safety. It is important to note that this does not mean that the same risk would necessarily be also judged preferable by an AV user, by a pedestrian or even by another AV developer. Risk perception varies between individuals and between segments of society, and is informed by risk appetite and underlying ethical principles.

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

$$\textit{Risk} = \textit{harm} * \textit{likelihood}$$

If an AV developer wants to differentiate between two risks which may be of equivalent magnitude, and further to trade one of these risks off against another, it is necessary to have greater transparency into this calculation. In particular, the developer

must be able to provide sufficient information to justify why these risks are different, and to demonstrate how his or her underlying ethical imperatives have informed trade-offs between different risks.

With this in mind, we propose the following factors that characterise risks beyond their magnitude (severity x likelihood). We treat the factors below differently to how we treat severity and likelihood, in that we do not impose a ranking on them, or hold an implicit brief that some values of these factors are "worse" than others in terms of ethics. The factors are informational only, and serve to differentiate one risk from another of equivalent magnitude. These factors are:

a) The exposed population

We propose identifying several broad categories for the population who might be exposed to any given risk. These will vary depending on the risk in question. For example, the population exposed to the risk of collision includes the AV passengers and other nearby road users, while the population exposed to the risk of emissions includes pedestrians and those living in nearby houses. Categories of stakeholders should include (not all will be relevant for all risks):

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

b) Effective risk horizon

This refers to the physical and temporal area in which the harm from a given risk is experienced. In physical terms, the harm may manifest close to the AV or further away, while in temporal terms the harm may manifest straight away or several years

later. We emphasise that a change in effective risk horizon does not necessarily affect the magnitude of the risk: that is, a "close" risk is not necessarily worse than a "distant" risk. Nevertheless, an AV developer may look to differentiate two equivalent risks by their effective risk horizons. Different risk horizons may cause the quality of the risks to be perceived differently, and may affect the way in which their associated economic, reputational or emotional effect is viewed. These will all have implications for the ethical imperatives of the AV developer. Some examples of risks which differ in terms of effective risk horizon (as well as potentially in magnitude) are:

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

c) Causes

Risk cause is also relevant to AV developers seeking to differentiate between equivalent risks. Although, as with the above factors, the cause of a risk does not necessarily affect the magnitude of that risk, there may be ethical, emotional and intuitive reasons why an AV developer might prioritise the mitigation of risks from particular causes. For example, Section 2.2 identifies regret avoidance as a

consideration in some risk decisions; a developer may choose to reduce risks from a cause which is also likely to engender regret.

4.2.1 Risk profiles and risk balancing

Previous work (Menon et. al., 2013) introduces the concept of risk profiles, each of which represents an approach to balancing and trading off risks. The risk profiles defined in previous work focus primarily on the magnitude of risk, and the justification for accepting an increase in one risk in return for a decrease in another. Two fundamental risk profiles examined in (Menon et. al., 2013) are the *fairness in improvement* profile (where an attempt is made to reduce all risks by the same magnitude) and the *fairness in outcome* profile (where an attempt is made to reduce all risks to the same magnitude). The authors also identify a number of confounding factors which can limit the practical application of these risk profiles, including the challenge of fully characterising two equivalent – but different – risks of the same magnitude.

The risk refinement factors identified above represent an approach that can be developed to address this challenge. Characterising a risk by its exposed population, effective risk horizon and cause allows more complex ethical arguments to be made about risk prioritisation and permits, for example, the ethical arguments that might be used in accepting a brief short-term increase in risk for a longer-term decrease in risk.

Constructing representative risk profiles from the risk refinement factors is beyond the current scope of this paper, and hence we restrict ourselves to noting the utility of such risk profiles in explicitly identifying and justifying the ethical imperatives behind risk trade-offs.

5. A process framework and ethics assurance case pattern

In this section we present a framework which enables the translation of ethical

principles into specific safety / design requirements, and also provides a tool for demonstrating the satisfaction of these requirements. The framework is in two parts: an ethics assurance case template pattern and a process framework for instantiating this template to produce an ethics assurance case. The process for instantiating the ethics assurance case template pattern uses the risk balancing and trade-offs (risk profiles) of Section 4, and must be embedded within a wider engineering lifecycle.

We first provide some background on structured assurance cases, and their history of use. We then build on this to define the template pattern and the process framework, including a diagram and textual discussion of each.

5.1 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 scrutinized (e.g. by a regulator) and defended. The adequacy of the argument is critical, and assurance cases typically build on a set of principles which must be satisfied within the overall argument construction. These principles define an assurance case *template pattern*, which is a recommended method of constructing the argument in order to minimise the chances of

introducing a logical, technical or semantic error. An assurance case template pattern therefore consists of a template argument built upon these principles. In practice, the template argument might contain a number of claims in order to support each principle, and each claim might also support multiple principles.

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. Consequently, argument template patterns for assurance cases to be used in the defence environment typically contain claim(s) corresponding to each principle:

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

Argument template patterns must be instantiated for use with a particular system. It is at this stage that specific requirements, specifications, V&V artefacts and other items of evidence are included to expand and support the claims.

A significant body of work already exists around the construction and instantiation of assurance case template 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).

5.2. Principles for ethics assurance case template pattern

In order to construct an ethics assurance case template pattern, we build on the four key

principles for assurance cases in the defence domain (MOD, 2017). We define four extended principles for the ethics domain 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. Process and assurance case framework

5.3.1 Ethics assurance case template

Figure 1 shows the ethics assurance case template pattern using Goal Structuring Notation (Assurance Case Working Group, 2018). We discuss this textually below.

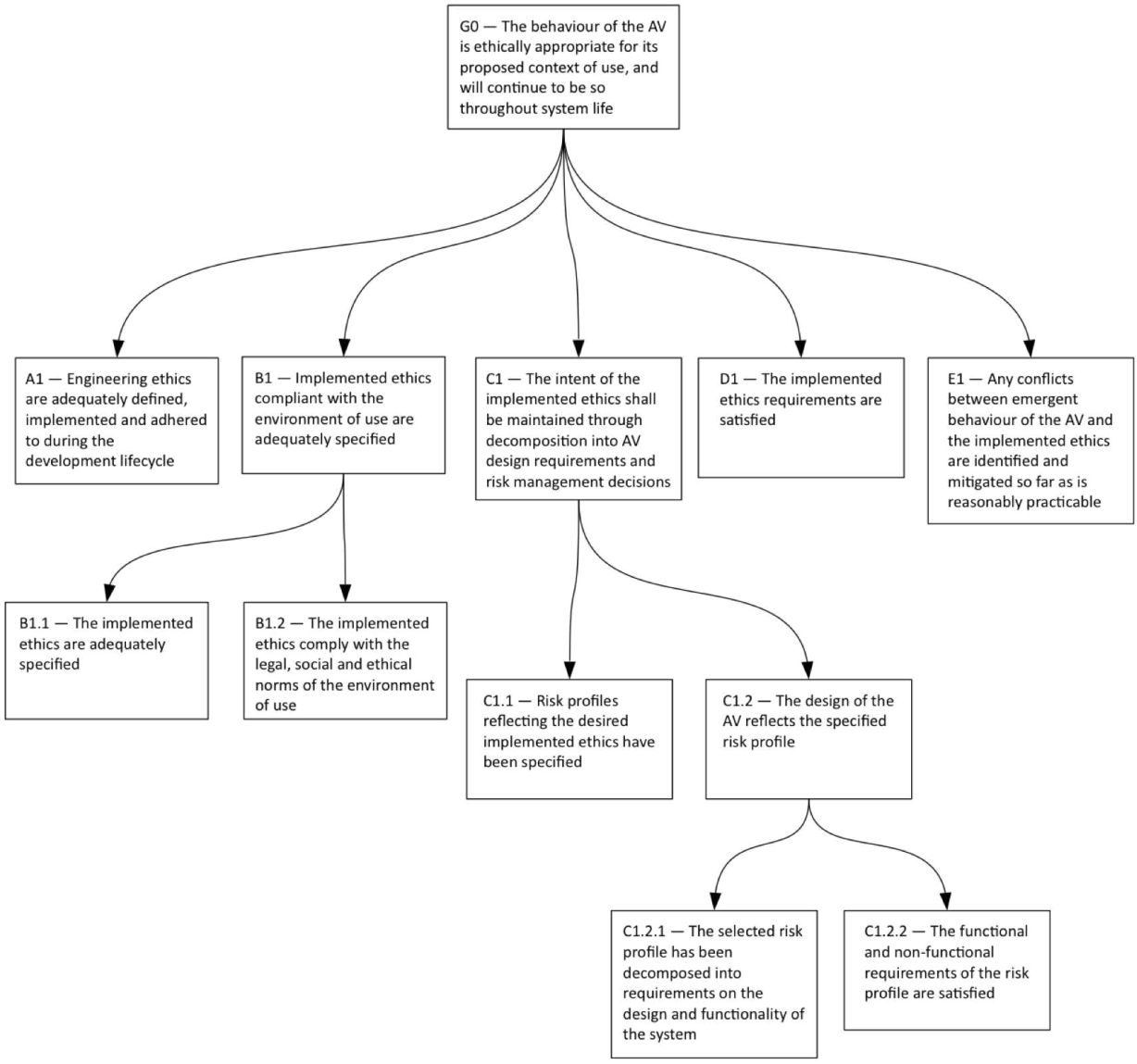

Figure 1: ethical assurance template argument pattern

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 human driver (IEEE Global, 2018), but rather that the AV should act in a way that a human might plausibly expect from an AV.

Claim C1: *The intent of the implemented ethics shall be maintained throughout decomposition into AV 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: *The risk balancing and trade-offs resulting from the implemented ethics across different environments have been specified in the form of a risk profile*

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 provision of an explicit description of the risk trade-offs and balances that have been performed, in the form of a risk profile.

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

5.3.1 Engineering-focused process

In order to be applicable to a given system, the ethics assurance case template that we describe above must be instantiated for that system. This is achieved via the second part of our framework: an engineering-focused process that uses the risk profiles of Section 4 to makes risk trade-offs and risk balances explicit. This process sits within the wider engineering lifecycle, and relies on artefacts from that lifecycle, such as test results, requirements specifications etc.

The process consists of several phases, as shown in Figure 2, with decision points and feedback loops between the phases. We emphasise that the decision points and phases

of Figure 2 relate only to the instantiation of the ethics assurance template. Consequently, Figure 2 deliberately excludes decision points and feedback loops which relate to the wider engineering lifecycle, such as feedback loops to indicate iteration over engineering requirements, or decision points to verify that such requirements are fulfilled. This ensures that instantiation of the ethics assurance template does not depend on a specific engineering lifecycle process being used.

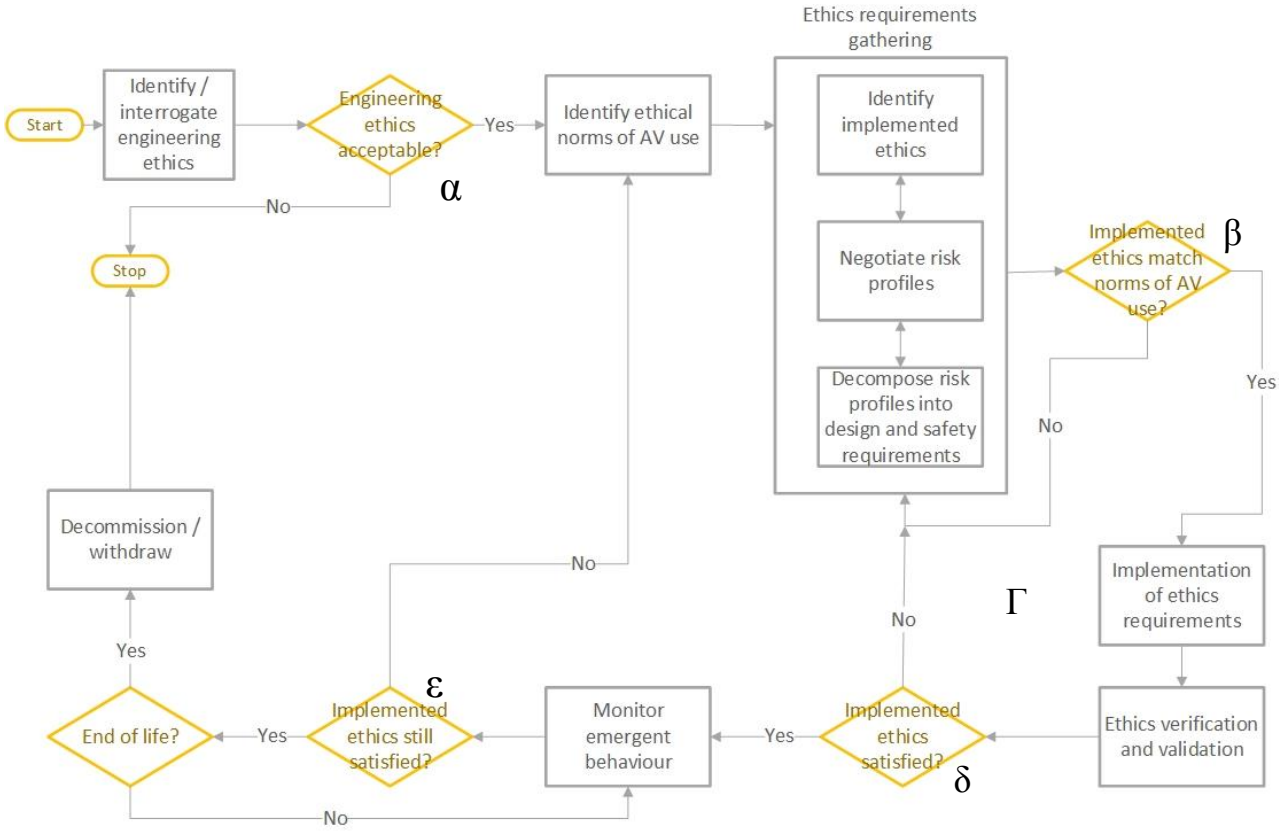

Figure 2: engineering-focused process

Phase 1: identify and interrogate engineering ethics

This phase instantiates Claim A1 of Figure 1, and consists of activities to identify and interrogate the engineering ethics of the developers. These ethical principles and codes of conduct are crucial to maintain the integrity of the AV development for stakeholders to have confidence in all artefacts produced during development. Because of the

importance of the developers' engineering ethics, concerns at this point may result in a hard stop for the project via decision point α.

Phase 2: Identify ethical norms of AV use

This phase partially instantiates Claim B1 of Figure 1 and consists of activities to identify the legal, social and ethical norms of the proposed AV use. These norms will be situation-dependent (an AV operating on a motorway will be subject to different norms than an AV operating near a school). As discussed earlier, this phase is not about determining how human drivers act in a given environment, but rather about determining how humans will expect the prospective AV to act in its environment of use situation.

Phase 3: Ethics requirements gathering

This phase instantiates Claim C2, and also completes the instantiation of Claim B1. It is a tri-part phase which also serves as an interaction and feed-in point for the requirements phase of the wider engineering lifecycle. During this phase the developers identify the implemented ethics (i.e. the ethics that will govern the behaviour of the AV), negotiate risk profiles which reflect those ethics and their associated risk trade-offs, and decompose these risk profiles into design and safety requirements. (There will of course, be additional requirements gathering activities defined within the engineering lifecycle, but these are excluded from the scope of Figure 2). Should the implemented ethics not match the norms of AV use in the proposed environment, this phase is iterated over again, as shown via the decision point β and its associated feedback loop in Figure 2. We note that in practice that gathering and refinement of ethics requirements is likely to be an ongoing process, occurring simultaneously with implementation of ethics requirements and verification and validation of ethics requirements. This iteration and simultaneous development is shown via the feedback loop Γ of Figure 2.

Phase 4: Implementation of ethics requirements

This phase serves as an interaction and feed-in point for the system development and implementation phase(s) of the wider engineering lifecycle. This phase incorporates development and implementation against the ethics requirements of the AV specification. There is no specific output from this phase in terms of instantiation of the ethics assurance case template.

Phase 5: Ethics verification & validation

This phase instantiates Claim D1 of Figure 1. It also serves as an interaction and feed-in point for the verification and validation performed during the wider engineering lifecycle. This phase incorporates verification and validation of the implemented AV system against its ethical requirements. Should this V&V identify that the ethical requirements have not been met (via decision point δ), there will be a return to Phase 3 to iterate over the gathering, refinement and subsequent implementation of these ethical requirements.

Phase 6: Monitor emergent behaviour

This phase instantiates Claim E1 of Figure 1, and consists of activities to monitor any behaviour of the AV which emerges after its deployment. Emergent behaviour may be due to the use of continuous machine learning, which requires the AV to adapt and refine its behaviours in response to new input. Similarly, emergent behaviour may be due to an original under-specification of behaviours, or over-restrictive assumptions about the environment in which the AV may eventually need to operate. Where any emergent behaviour is identified, a further assessment activity must be made to determine whether this behaviour complies with the ethical requirements of the AV. Because such emergent behaviour may relate to a new or unforeseen use of the AV, or environment of use, any contravention of the ethical requirements results in a return to

Phase 2 (via decision point ε) in order to iterate and expand upon the norms of AV use for this new environment.

Phase 7: decommission and withdrawal

This phase corresponds to the end-of-life point for an AV system. End-of-life decisions may be taken where the system is superseded by another, or where legal, ethical, technological or societal factors lead to a situation where withdrawal of the system is recommended. There is no specific output from this phase in terms of instantiation of the ethics assurance case template.

6 Conclusions

In this paper, we have discussed the ethical concerns around the introduction and use of AVs, focusing on the perceived risk these present, and the acceptability of this risk. We have identified a number of factors which affect risk perception and risk acceptance, particularly where ethical principles such as altruism or "fairness" are concerned.

One of the most significant underlying ethical issues we have discussed is the situation where different risks are balanced against each other, with an increase in one risk being traded off against a decrease in another. This is of particular concern to AVs where risk is transferred between different segments of society (e.g. from the AV passenger to pedestrians or other road users). At present it is not always clear when such risk trade-offs have been made by developers, or what the ethical factors motivating these are. We argue that this lack of clarity can lead to a lack of informed societal consent to these risks.

In order to increase transparency of the ethical factors which drive such risk decisions, we have identified a methodology for explicitly translating the underlying ethics of AV behaviour into safety and design requirements on the AV. This methodology makes use of two components: an ethics assurance case template, and an

engineering-focused process for instantiating this template. Together, these ensure that the underlying ethical and risk-balancing choices of the AV developers are identified and explicitly justified. This is achieved via the use of risk profiles, which require developers to describe the approach to risk reduction that they have taken, and to justify their risk transfer decisions.

The framework we propose is also applicable to fully-autonomous systems in other domains with ethical concerns, such as the military and healthcare fields. We propose to expand this work in future to apply this methodology to a working case study, either in the automotive, military or healthcare domains. A foundational step in this will be to build a taxonomy of risk profiles in conjunction with industry personnel. This will allow us to represent the risk transfer decisions which are made in a real-world scenario.

We also propose to extend the framework to address the issue of confidence. Like safety, ethics is a limit concept (Kelly, Habli, Nicholson, Megone & Mcnish, 2014) and the degree of confidence stakeholders have in the ethics of a system will be dependent on the evidence they have been shown. We propose further study into the societal perception of AVs in order to identify what factors increase confidence in the ethical principles governing the AV, and what steps can be taken to mitigate a lack of confidence. As part of this work we will also consider ways to formalise certain ethical principles relevant to AV introduction and operation, including the principle of double effect and Kantian ethics, along the lines of (Bentzen, 2015; Linder & Bentzen, 2018).

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, 503-546

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.

Baron, J. (1985) *Rationality and intelligence*. New York, Cambridge University Press.

Bell, D. (1983). Risk premiums for decision regret. *Management Science* Vol 29(10), 1156—1166.

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.

Cokey, E. & Kelley, C. (2009) Cognitive Abilities and Superior Decision Making. In *Judgement and Decision Making*, Vol 4(1), 20 – 33.

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.

Evans, J. (1989). *Bias in human reasoning: Causes and consequences*. Brighton, Erlbaum

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

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

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263-291.

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/.

Von Neumann, J. & Morgernstern, O. (1947). *Theory of Games and Economic Behaviour.* Princeton University Press.

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.

Reyna, V. (2004). How People Make Decisions That Involve Risk: A Dual Process Approach. Current Directions in Psychological Science, Vol 13(2), 60 – 66.

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.

Somasundaram, J. & Diecidue, E. (2015). Regret Theory and Risk Attitudes. Journal of Risk and Uncertainty, Issue 2-3, pp. 147-175

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.

Wang, S. & Zhao, J. (2019). *Risk Preference and Adoption of Autonomous Vehicles*, Transportation Research A.

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.