Ethics and the safety of autonomous systems

Catherine Menon*, Rob Alexander†

University of Hertfordshire*, University of York†

Abstract *The ethical landscape surrounding the introduction of autonomous vehicles is complex, and there are real concerns over whether the operational safety of these systems can be adequately demonstrated. In this paper we focus on the ethical factors relevant to the design and safety justification of autonomous systems, considering issues such as risk transfer, ALARP considerations, capability vs risk trade-offs and emergent behaviours. We look beyond the "trolley problem" to consider how design decisions can reflect a wider ethical framework. We also look at the wider landscape around the emergence of autonomous systems, with a particular focus on the driving social factors which encourage early adoption of new technologies in this domain. We present some arguments for encouraging an explicit discussion of social and ethical factors within the safety framework for autonomous systems.*

1 Introduction

Autonomous systems (AS) have been proposed for use in multiple domains, including nuclear, medical, defence, rail, maritime and automotive. The ethical requirements across each of these domains will inevitably differ, and in many cases there is no consensus as to which system behaviours would be deemed ethically appropriate. The SCSC Safety of Autonomous Systems Working Group (SASWG) has been engaged in producing guidance on the safety of autonomous systems, and this paper constitutes a summary of our current position on ethics.

We note that ethics considerations do not relate solely to safety, and a discussion of AS ethics may include environmental impact, economics, manufacturing processes and adequate financial investment. While we consider these as influencing factors, detailed analysis of these aspects is beyond the scope of the SCSC SASWG and therefore of this paper.

2 Ethical background

The first ethical question we introduce relates to the position of humans in the decision-making process. When we move from human actors to automated ones, we move the intelligence in the decision from conditions which are potentially subject to extreme time stress to a much calmer, slower-paced environment. That is, the way in which the AS reacts is determined by the programming and algorithm decisions, made by developers during the design and implementation stages. (This excludes the possibility of AS which use dynamic machine learning, as such systems may reasonably be assumed to be infeasible in the short-term future).

This may raise the standard of ethical performance the public expects. For example, in the case of a human driver, any decisions made in a collision situation are judged according to that environment and drivers – except where their actions have been negligent – are generally not considered culpable should they make the “wrong” decision (Lin 2015). This is also seen in the military domain with regards to the rules of engagement, and discussed further in Section 2.3. An engineer developing an AS, by contrast, is not under such pressure, and may therefore be expected to ensure that the AS reacts in a societally acceptable way, regardless of how a human actor might.

More generally, there is the question of risk acceptability. It is not clear that the general public will necessarily be willing to accept the same risk when it is posed by a machine as opposed to a person. To some stakeholders, ASs may be acceptable only when they represent a significant decrease in risk compared to human actors. Consent to the risk presented by a system, and the extent to which this risk can be justified, is discussed in more detail in Section 3. In order to provide examples for this, we present a comparison below of the primary ethical concerns in some of the identified domains. This is not intended to be an exhaustive discussion of ethics in each of these fields, for which the interested reader is referred to existing literature.

2.1 Automotive

The automotive domain is one in which the ethical aspects can be perhaps most readily characterised, with the primary concern being framed in terms of the “trolley problem”. This refers to a thought experiment in which a train / trolley is on a set of tracks which will cause it to collide with a number of people. The observer is asked whether s/he would choose to switch the train to a second set of tracks which will cause it to collide with a single person only. Amendments and extensions to the trolley problem have couched the problem in terms of an active vs. passive choice as well as experimented with the relative “worth” of each person affected.

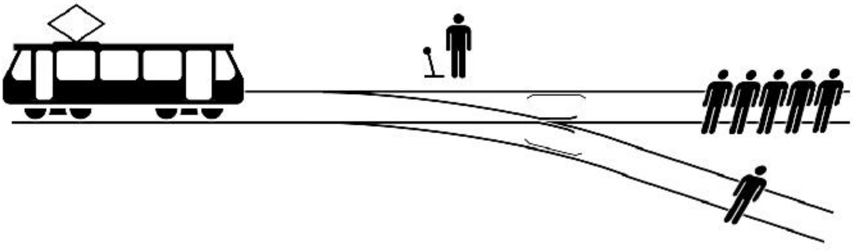

Fig. 1. The trolley problem (McGeddon, 2016)

The automotive domain also presents some further ethical questions. In (Lin, 2015) a case is discussed whereby an autonomous vehicle (AV) may choose to position itself within a lane closer to a smaller car than to a truck. This decision might be justified in two ways: firstly, that this behaviour is typical of a human driver, and secondly that this reduces the risk to the AV. From a safety perspective, this decision has prioritised the safety of the AV occupants – and the truck occupants – over that of the smaller car. Such a decision would need to be justified within the safety case and from an ethical perspective.

Alternatively, the AV may take the opposite course; choosing to drive closer to, or to impact, a heavier vehicle or a vehicle with safety systems known to be better. In this case the severity of an accident may be reduced, compared to an impact with a vehicle with poor safety systems. However, implementing such a decision into the behaviour of the AV represents a deliberate choice to increase the risk to drivers of certain vehicles and must be justified ethically. Other situations discussed in the existing literature include the decision of an AV to sacrifice itself (place itself in the path of another vehicle to save a third party from impact), as well as choosing to impact a motorcyclist wearing a helmet over one not wearing such protective devices (Gerdes and Thornton 2016).

The automotive domain, among others, is also subject to commercial pressures. There is significant public interest in self-driving cars, and engineering companies are alert to the advantage of bringing out the “first of kind” of an AV. The high-profile nature of commercial AVs can encourage the categorisation of safety as a competitive advantage. This means that best practice can be difficult to establish, and known problems may not be shared for reasons of commercial interest.

In addition, there are currently no applicable standards which fully address the safety of AVs, including safety of the intended function (ISO 26262 2011). Consequently, while there is a clear economic and reputational imperative for a company to bring out the “first of kind” in autonomous vehicles, it is much less clear that such an AV could be demonstrated to be acceptably safe. There is a risk that the push to produce and market AVs can encourage “quick and dirty” practices during the development lifecycle which can have an effect on the system as released to the public. While standards do exist around ethical design of systems

(IEEE Global Initiative 2016), these are relatively new and their general applicability has not been fully determined.

2.2 Medical

There are a significant number of ethical concerns in the medical domain, which we will not attempt to discuss exhaustively here. Unlike the automotive domain, ethics in the medical domain typically do not involve trolley problems. Rather, medical ethics problems tend to relate to the trading off of risk for medical benefit, or Quality Adjusted Life Year (QALY). While these decisions are not specific to ASs, a question remains around the ability of such systems to adequately judge the quality of life, as well as the willingness of the public to accept such a decision made by a machine.

In the medical domain there is also a cost associated with not delivering care, for example where the remoteness of a region or lack of available SQEP caregivers means healthcare access is limited. In these situations the introduction of ASs to perform diagnostic functionality or allow minimally invasive surgery may not replace an existing human capability. This differs from the automotive domain, where the transport capability is generally understood to be already in place (i.e. human drivers). An ethical question in the medical domain therefore might concern whether the risk posed by an AS is justifiable in the absence of any existing capability for diagnosis and treatment. To a lesser extent this is also the question when considering assistive “companion” AS technologies; the main role for these may be in situations where no human caregivers are available.

Privacy is also a significant ethical question in the medical domain. While risks may be reduced by sharing medical information (e.g. with other systems, with healthcare practitioners), this must be balanced against the privacy requirements of an individual.

Further ethical complexities include the possibility of skills degradation (whereby human actors lose their diagnostic and treatment skills), diagnostic capabilities and side-effects. In particular, ethical complexities around autonomous diagnostic devices involve the possibility of false negatives (resulting in treatment being erroneously withheld) and false positives (treatment will be given unnecessarily). While false negatives have an obvious safety impact, false positives can trigger medical intervention or further potentially harmful diagnostic tests.

2.3 Defence

In the defence domain, there are a number of ethical concerns relating to the acceptability of ASs. One of these relates to the ability of such systems to make a firing decision, with armed sentry systems being an example of these. Ethical concerns may include whether it is justifiable to fire on friendly troops, the impact of mistaken identity and the conditions under which civilian casualties are acceptable (Lin 2008).

Even where systems do not autonomously make target engagement decisions, there is an ethical concern over such systems as a replacement for humans in combat situations. It has been argued that human operators of UAVs may be more willing to engage targets, because of a distancing effect due to the geographical distance between them (Borenstein 2008) This presents an ethical disincentive to the introduction of such systems. However, by contrast, ASs would not be subject to the extreme stress human soldiers are placed under, and so may be less likely to contravene rules of engagement (Lin 2008).

On a larger scale, it has also been suggested that the use of military AS systems subverts casualty aversion (Walsh and Schulzke 2015). Casualty aversion is an ethical constraint resulting from public reluctance to support a given military action due to the human cost. In this way the use of such ASs can be seen to create a moral hazard, in removing the risk associated with their actions. One consequence of this may be that the introduction of ASs in the defence domain could potentially lead to an increase in conflict frequency due to public willingness to encourage this. However, this could also lead to an increased willingness to undertake conflict for humanitarian reasons.

2.4 Systems of ethics

While we do not attempt to provide an exhaustive background to ethical philosophies, the following ethical theories provide relevant terminology.

2.4.1 Consequentialism

Consequentialism (Goodall 2014) is an ethical theory which prioritises outcomes: consequentialist ethics deems acts to be morally acceptable if they lead to a good outcome. This is sometimes summarised as “the end justifies the means”. A consequentialist approach to AS safety would be to seek to reduce overall harm (e.g. by minimising the number of people harmed; a consequentialist solution to the trolley problem would be to switch the trolley onto the track with a single person). Consequentialism as an ethical theory is aligned with more general safety criteria (Health and Safety Executive 2001). in terms of minimising harm, but does not take into account questions of risk responsibility, informed consent for acceptance of risk and calculations relating to acceptable exposure due to work.

2.4.2 Deontological

By contrast, deontological theories of ethics prioritise acting in accordance with explicitly stated duties and rules (Goodall 2014). Deontological ethics therefore does not require the AS to consider the outcomes, but merely to act in accordance with pre-programmed rules (which may include, for example, a rule that the AS must not injure – or cause to be injured – any person). While encoding such rules is conceptually simpler than requiring the AS to perform calculations minimising harm, deontological ethics does require the identification of rules for every situation the AS may find itself in. A deontological approach to the trolley problem would be to consider whether rules exist which govern the acceptability of switching the trolley to a different track, regardless of the risk exposure to any individuals.

2.4.3 Virtue ethics

A third ethical imperative relevant to ASs is the concept of virtue ethics, typically presented in terms of self-sacrifice (Lin 2015). This discusses the extent to which an AS should choose to act altruistically, according to some stated definition of this. An automotive AS adhering to virtue ethics would choose to sacrifice itself and its passenger in order to reduce harm to a third party, while a military AS adhering to virtue ethics would potentially be of questionable utility.

3 Ethics and risk

One of the fundamental issues identified in all domains at the intersection of ethics and safety is the question of risk transfer. Although it may be possible to argue that the introduction of ASs in certain situations reduces the overall harm, from an ethical perspective this may not be sufficient. This is because if risk is transferred away from some exposed people and placed on others then the transfer must be explicitly justified, even where the overall risk is reduced.

In more detail, it may be the case that introduction of ASs causes a segment of the population to face either an absolute or a relative increase in the proportion of risk which they bear. An example of this can be seen in the trolley problem: consider an AS which causes fewer collisions, but uniformly chooses to impact smaller pedestrians when a collision is unavoidable. Such an AS would cause these pedestrians to bear a disproportionate amount of the overall risk, when compared to a human driver. The medical domain also provides examples of this, most notably compulsory vaccination programs. These reduce overall mortality by promoting herd immunity, but a small number of vulnerable individuals are harmed by the vaccine. In this example, overall risk has been reduced, but the individuals in question are exposed to an increase in the proportion of that risk which they bear.

The question of consent is also relevant, in that exposed parties have not necessarily consented to bearing the portion of risk allocated to them. In the military domain, affected civilians may not even be aware that ASs are in use, while in the medical, automotive and rail domains it is more likely to be the case that exposed parties are unaware of the principles governing AS behaviour. It is arguable that in some domains, such as automotive, an explanation of the ethical behaviour of the AS should form part of the product certification.

A further concern is the impact of ASs when considered from a systems of systems (SoS) perspective. Whether the wider system relates to rail, to the road network, to a patient’s overall medical care or to defence capability, ASs comprise only one component within an interconnected system. Potential interactions with other ASs must be considered, as well as interactions with infrastructure, human operators and third parties. Particularly where ASs make use of machine learning algorithms, there is the potential for unforeseen interactions and emergent behaviour. For example, in the automotive domain we may see an increase in traffic jams due to all AVs following the same route, as it is in the interest of no individual AV to change route. Local optimisations made by ASs in the medical domain may cause patients to be sent for unnecessary scans and treatments, while in the defence domain automating the task of learning which targets are acceptable – and basing target engagement decisions off these – is likely to present significant concerns. More generally, there may be an issue if different ethical imperatives are embedded within different systems which interact. It may not be possible for these systems to coexist in an ethically compatible manner.

3.1 Risk balancing and risk transfer

If risk transfer is understood to be the foundation for these concerns then the ethical problems can be rephrased in terms of the trade-off associated with reducing one risk posed by an AS at the potential cost of increasing another risk. This gives us the ability to discuss ethics – at least partially – in the language of safety.

In general, for any given (autonomous or non-autonomous) system there may be multiple ways to reduce the overall risk posed to As Low As Reasonably Practicable (ALARP). Individual risks can be traded-off, or balanced against each other, where an increase in one risk is accepted in return for a decrease in another. This concept is discussed in standards primarily within the nuclear domain (HSE 2006) (ONR 2013), which emphasise the need to balance individual risks within a system and consider established good practice. However, outside this domain many safety guidance documents (HSE 2001) provide little information on how to balance risks and make these choices, requiring only that the overall system risk should be ALARP.

Risk trade-offs and balances can happen at three levels throughout system development. At the development level these are relatively common, as many development choices imply that a decision must be taken between the risks associated with each possible approach. For example, choosing to develop software in C instead of SPARK ADA may provide increased access to experienced developers, but at the cost of static analysability.

At the system level, as discussed, one risk posed by the system may be mitigated at the cost of potentially increasing another. Finally, at the external level, an increased safety risk may be associated with a benefit in another domain, such as security. This is discussed further in Section 4.5.

4 Risk Profiles

In (Menon et al. 2013) we presented a number of different risk reduction approaches, or *risk profiles*, which provide alternative ways of balancing individual risks in order to achieve an ALARP system risk. An ontology of these is briefly given below. It is unlikely that a single risk profile will be suitable for balancing all system risks, and therefore we would recommend that these profiles be combined and customised as needed.

4.2 Fairness in improvement

The aim of this approach is to achieve a similar absolute risk reduction for all individual risks. A fairness in improvement approach prioritises the reduction of all risks *A*, *B*… *N* regardless of the relative cost of these reductions (provided these are reasonably practicable), and regardless of whether making these reductions to one risk A means that for technical reasons further reductions cannot then be made to another risk B. Using a fairness in improvement approach can mean that no individual risk is as low as technically possible when considered in isolation. However, this approach ensures that the risk reduction effort confers a certain minimum benefit on all system risks.

A fairness in improvement approach for AS risk reduction may correspond to attempting to mimic the actions and risk reduction behaviour exhibited by a human actor. The risks posed by the AS will therefore bear a similar proportionate relationship to each other as the risks posed by a human actor, although the overall system risk may be lower. In the military and automotive domains, this may correspond to emphasising the need for AS functionality to match human behaviour (e.g. in a trolley problem scenario, or when making firing decisions). In the situations encountered in the medical domain where no comparable human actor is available, this would require balancing the risks associated with incorrect diagnoses or surgery against the risks associated with a lack of treatment.

4.3 Fairness in outcome

The aim of this approach is to achieve a similar level of risk for all individual risks. Fairness in outcome means that our risk reduction attempts prioritise the reduction of a more severe risk A over the reduction of a less severe risk B. This is not affected by the relative cost of reducing risks A and B compared to each other, or whether making these reductions to A means that for technical reasons further reductions cannot be made to B. Using a fairness in outcome approach can imply that the risk reduction efforts are concentrated on only a few risks, with no benefit for the other risks. However, a benefit of this approach ensures that the areas of greatest risk are targeted by reduction efforts.

A fairness in outcome approach for AS risk reduction may correspond to a focus on reducing the greatest risks posed by the AS. In the case of the automotive domain this presents a solution to some manifestations of trolley problem: impact with other vehicles is likely, for example, to be a preferred hazard over impact with pedestrians. It is worth noting that Google have adopted a partial fairness in outcome approach, stating that their priority is to avoid impacting unprotected road users (Automotive IQ 2017).

4.4 Long-term risk benefit

The question of system risks that change over time can also be relevant when balancing individual risks. Standards such as (HSE 2006) also consider the possibility of accepting a higher short-term risk if this results in a long-term risk reduction.

For AS risk reduction, taking a long-term risk benefit approach prioritises the introduction of ASs, along with any concomitant short-term increase in risk, should it be possible to demonstrate that this would lead to fewer lives being lost over the long-term. Long-term risk benefit requires explicit justification within the safety case, as it may not be possible to demonstrate that in the short term the system risk is ALARP. Consequently, long-term risk benefit should be used only to customise and refine other risk profiles.

4.5 External risk transfer

Risk transfer refers to the situation where multiple components or subsystems interact, such as within a SoS. In this case, an ALARP claim for each subsystem considered in isolation does not necessarily lead to the lowest overall system risk. In these situations an increase in a local risk associated with one system may be accepted in return for a decrease in the risk associated with the wider system. This is presented in further detail in (Menon et al. 2013).

More generally, in some cases an increase in a safety risk may result in a benefit in an external domain. For example, the presence of certain security features such as Intrusion Detection Systems (IDS) provides a security advantage while making it harder to demonstrate the safety of the system (amongst other concerns, IDS need to be regularly updated, which is difficult given the rigorous testing and validation required by safety-critical systems (Johnson 2014)). It should be noted, however, that this external risk transfer cannot be deemed acceptable from an ALARP perspective, as the ALARP principle does not consider benefits outside the safety domain.

5 Safety, ethics and development

Risk profiles allow us to bring safety and ethics together for AS behaviour by making explicit the risk balancing and trade-offs inherent in any ethical decision. It will also be necessary to justify these decisions, both from an ethical and a safety perspective. In order for all stakeholders to adequately understand the implications of these decisions we propose that the argument be presented within an explicit “ethics case”, comparable to – and cross-referencing – the safety case. In this section we discuss how such a case may be constructed.

5.1 Engineering and implemented ethics

When referring to ASs development and operation there are two interrelated but distinct applications of ethics and ethical systems. The first of these we will term *engineering ethics* and the second *implemented ethics* (or machine ethics, in AI terminology).

Engineering ethics refers to the ethical principles adhered to by engineers during system and software development. These may be in the form of principles or codes of conduct formalised by a professional organisation (RAEng 2017). They typically include criteria such as honesty, integrity, respect for law and the public interest, accuracy, rigour, fairness, objectivity and leadership. In addition, they encourage further thought and assessment to determine if any given engineering action is ethically defensible. It is important to note that adherence to a code of engineering ethics does not, in itself, mean that the behaviour of any resultant system will necessarily be considered ethical by all stakeholders. However, adherence to a code of engineering ethics helps to support arguments about the behaviour and properties of the system by providing confidence in the integrity of any lifecycle artefacts. Should developers not adhere to any professional code of ethics, any argument about the safety of the system or its behaviour can only be weakly supported.

Implemented ethics (or machine ethics), by contrast, refer to the ethics which govern the behaviour of the AS itself. These include deciding whether to prioritise the safety of the AS and its operator over third parties, deciding what functionality to deploy in given situations (e.g. target engagement decisions), deciding which of multiple third parties to prioritise where harm is inevitable, as well as making decisions related to the balance between safety, security, privacy and trust.

Unlike engineering ethics, there may not be consensus on what the “right” implemented ethics are. Acceptable ethical behaviour will vary across different societies (including different countries) as well as different domains of use.

6 Ethics case and argumentation

As with safety arguments, there is no single method of creating a failsafe argument to support claims relating to AS ethics. However, there do exist some generalised ethical foundations (IEEE Global Initiative 2016) relevant to all aspects of an AS.

In order to reflect our focus on ethics affecting safety, we propose the following principles to demonstrate the ethical integrity of the system. These echo the principles governing the integrity of Programmable Electronics (PE) in Annex D of (MOD 2015), and are aligned with the ethical foundations of (IEEE Global Initiative 2016).

1. **Principle P1:** Ethics requirements governing the AS behaviour shall be defined.
2. **Principle P2:** The intent of the ethics requirements shall be maintained throughout decomposition.
3. **Principle P3:** Ethics requirements shall be satisfied.
4. **Principle P4**: Any AS behaviours which conflict with the ethics requirements (“ethically hazardous” behaviours) shall be identified and mitigated.

(MOD 2015) defines one further principle relating to the confidence which has been achieved in addressing the PE safety principles. An analogy in the ethics domain would be the definition of an ethics proportionality principle and recommendations as to how this may be achieved or demonstrated. This is at present beyond the scope of this work.

We present a method of incorporating these principles into an ethics case argument, which aligns with the ethical foundations identified in (IEEE Global Initiative) as well as relevant safety and legal criteria (HSE 2001). The overall claim is:

G0: The behaviour of the AS is ethically appropriate for its proposed context of use.

This claim is supported by five sub-claims:

A1: Engineering ethics are adequately defined, implemented and adhered to during the development lifecycle.

B1: Implemented ethics are adequately specified and comply with the legal, social and ethical norms of the environment of use.

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

D1: Behavioural outcomes of the implemented ethics are satisfied.

E1: Any conflicts between the AS behaviour and the implemented ethics are identified and mitigated so far as is reasonably practicable

The following sections address each of these claims in further detail.

6.1 Claim A1

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

The purpose of this claim is to demonstrate that the engineering codes of practice and prescribed ethical principles are not compromised or impacted by any decisions relating to the ethical behaviour which it is decided the AS should demonstrate.

The desired engineering ethics may be identified by referencing codes of conduct, domain good practice and relevant previous decisions and their adequacy should be justified. Evidence to support this claim may be in the form of Continuing Professional Development records, audit records, lifecycle artefacts, documented processes and policies and so forth.

6.2 Claim B1

B1: *Implemented ethics are adequately specified, and comply with the legal, social and ethical norms of the environment of use.*

This claim fulfils principle P1, and for clarity of argument may be usefully broken down as shown in the following template example.

B2: The implemented ethics are adequately specified.

This specification may be in the form of references out to legal documents, to standards and policies, to previous system design decisions, records of public consultations and so forth. The specification of implemented ethics must be sufficient to address all issues raised in Section 5.1, as well as to provide a justification that the issues under discussion are sufficient and complete.

B3: The implemented ethics comply with the legal, social and ethical norms of the environment of use.

As stated in (IEEE Global Initiative 2016), the norms of the relevant community (or environment of use) must be considered when assessing the behaviour of the AS. The implemented ethics must be compatible with these norms. It should be noted that this does not mean that an AS should behave in exactly the same way as a human actor (that is, the implemented ethics do not have to be identical to the ethics currently embedded within the environment of use), but the two must be compatible, and any discrepancies identified and a justification provided.

6.3 Claim C1

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

This claim fulfils principles P2, and for clarity of argument may be usefully broken down as shown in the following template example.

C2: System design and AS functionality are adequately specified.

This sub-claim should be supported with evidence relating to the system design and implementation. Its intent is to demonstrate that the AS system design is specified sufficiently well enough to reduce the likelihood of unexpected behaviours. Should the intended behaviour or the design of the AS be underspecified, then it becomes much harder to predict whether the resultant operational actions of the AV will be considered ethically acceptable.

C3: Design decisions and risk management decisions are informed by the specified implemented ethics.

This claim fulfils Principle P2 and should be supported by nomination and definition of a specified risk profile (customised if required, as described in Section 4). It must also be demonstrated that this risk profile reflects the desired implemented ethics. The nomination of a risk profile, with the consequent requirement that describes a mechanism for reducing the system risk ALARP, is necessary in order to ensure that the specified implemented ethics do not contradict any of the legal requirements around safety (HSE 2001).

For example, should the implemented ethics require that the AS behaviour mimic the behaviour of a human actor (thereby resulting in no change in relative risk distribution across the wider system from the replacement of human actors with ASs), then we would expect to see a “fairness in improvement” risk profile selected. In practice, the desired implemented ethics are likely to be sufficiently complex such that a significant amount of customisation is needed to any of the risk profiles of Section 4.

Secondly, this claim should be supported with evidence that the risk management and risk reduction decisions reflect the selected risk profile. In practice, this may best be done by referring out to individual claims in the safety argument and demonstrating how the risk prioritisation decisions have been reflected in the mitigations.

6.4 Claim D1

Claim D1: *Behavioural outcomes of the implemented ethics are satisfied.*

This claim fulfils Principle P3. Satisfaction of it first requires the identification of what behaviours from the AS are required by the implemented ethics principles. These may not be immediately obvious and it is likely that some textual analysis of these ethics will need to be performed. Demonstrating that the AS performs such behaviours is likely to involve significant evidence in the form of system verification and validation, which may be cross-referenced from the safety case. Traceability between high and low level functional and non-functional requirements must also be demonstrated, as must traceability between these requirements and verification.

6.5 Claim E1

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

This claim fulfils Principle P4 and is accounted for by the fact that, like safety, ethics is a limit concept (Habli et al. 2015). Just as a system cannot be guaranteed to be absolutely safe, it cannot be guaranteed to be absolutely ethical (this is exacerbated by the difficulty in adequately specifying a comprehensive set of ethical principles).

This claim should therefore be supported by a gap analysis of the AS behaviour and the behaviour that would be expected according to the implemented ethics (claim B1). Any gaps – conflicts of the AS behaviour and the implemented ethics – may be thereby identified and efforts made to mitigate them. It is unlikely that the AS behaviour will be fully defined, and hence any gaps or conflicts may need to be derived from the functional and non-functional requirements and the implemented ethics. Equally, it is very likely that the implemented ethics will not exhaustively describe all possible behaviours of the AS; some of these may even have no ethical implications. For any identified conflicts (ethically hazardous behaviours), the argument must demonstrate that mitigations have been put in place to reduce the effect of these conflicts so far as is reasonably practicable. This parallels the ALARP requirement for safety, and similar argument techniques may be used.

7 Conclusions

In this paper we have identified the ethical landscape and imperatives that govern discussion of AS behaviour across multiple domains. We have introduced and formalised the concept of risk trade-offs, and considered the ethical drivers behind these. We have also identified the need for transparency in risk balances and risk trade-offs in order that consent from stakeholders may be obtained.

We have presented a methodology for arguing that the behaviour of an AS meets ethical criteria deemed relevant to safety. This methodology draws on aspects of safety argumentation to support a number of claims relating to the definition of ethically acceptable behaviour, the applicability of this in the proposed environment and the design decisions made during AS development. We draw on the concept of risk profiles to transform ethical principles into the language of safety and to provide a foundation for discussing how our ethical principles impact our risk mitigation decisions.

We distinguish between the principles of ethical conduct constraining the professional actions of engineers, and the principles of ethics constraining the behaviour of the systems these engineers design. We recognise that ethics of system behaviour, like safety, is a limit concept and extend the consideration of ALARP into the ethical domain. This allows us to examine whether the behaviour demonstrated by the AS is sufficiently close to the ethically desired behaviour in the environment of use.

There is the potential for significant further work in this area, particularly in the areas of balancing risk trade-offs. It would be of value to further extend the ontology of risk profiles to consider which refinements are of most use across multiple domains. There is scope for considering the extent to which safety, security, ethics, trust, legal and regulatory factors interact, and how the requirements of these can be balanced for a general autonomous system. In addition, there is currently an area of work relating to confidence in the satisfaction of ethics requirements, which may have an analogy to confidence in the satisfaction of safety requirements; further research in this area would go some way towards addressing this.

Acknowledgments The authors wish to thank the SCSC Safety of Autonomous Systems Working Group.

References

Lin, P (2015) Why Ethics Matter for Autonomous Cars, Autonomous Driving, Springer, pp 69 – 85.

McGeddon (2016) The Trolley Problem. https://commons.wikimedia.org/wiki/File:Trolley_problem.png, used under CC-BY-SA. Accessed 23 October 2017.

Gerdes, J, Thornton, S (2016) Implementable Ethics for Autonomous Vehicles, Autonomous Driving, Springer, pp 87 – 102.

International Standard for Organisation (2011) Road Vehicles – Functional Safety, ISO 26262.

IEEE Global Initiative (2016) Ethically Aligned Design, IEEE Standards v1.0.

Lin, P, Bekey, G, Abney, M (2008) Autonomous Military Robots: Risk, Ethics and Design, US Office of Naval Research.

Borenstein, J (2008) Ethics of Autonomous Military Robots, Studies in Ethics, Law and Technology, Vol. 2.

Walsh, J, Schulzke, M (2015) The Ethics of Drone Strikes: Does Reducing the Cost of Conflict Encourage War, United States Army War College Press.

Goodall, N (2014) Machine Ethics and Automated Vehicles, Road Vehicle Automation, pp 93 – 102.

HSE (2001) Reducing Risks, Protecting People, Health and Safety Executive.

Goodall, N (2014) Ethical Decision Making During Automated Vehicle Crashes, Transportation Research Record Journal of the Transportation Research Board, Vol. 2424.

HSE (2006) Safety Assessment Principles for Nuclear Facilities, Health and Safety Executive.

ONR (2013) Guidance on the Demonstration of ALARP, Office for Nuclear Regulation.

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

Automotive IQ (2017) Autonomous Vehicles and the Trolley Problem. https://www.automotive-iq.com/infographics/autonomous-vehicles-trolley-problem. Accessed 14 September 2017.

Johnson, C (2014) Barriers to the Use of Intrusion Detection Systems in Safety-Critical Applications, SAFECOMP, LNCS Vol. 9337.

RAEng (2017) Statement of Ethical Principles, Royal Academy of Engineering.

MOD (2015) Defence Standard 00-56 Part 1, UK Ministry of Defence, Issue 6.

Habli, I, Kelly, T, Macnish, K, Megone, C, Nicholson, M, Rae, A (2015) The Ethics of Acceptable Safety, Proceedings of the 23rd Safety-critical Systems Symposium.