Towards a Framework for Safety Assurance of Autonomous Systems

John McDermid, Yan Jia, Ibrahim Habli
Department of Computer Science, University of York, UK
john.mcdermid, yan.jia, ibrahim.habli @york.ac.uk

Abstract

Autonomous systems have the potential to provide great benefit to society. However, they also pose problems for safety assurance, whether fully autonomous or remotely operated (semi-autonomous). This paper discusses the challenges of safety assurance of autonomous systems and proposes a novel framework for safety assurance that, *inter alia*, uses machine learning to provide evidence for a system safety case and thus enables the safety case to be updated dynamically as system behaviour evolves.

1 Introduction

This paper addresses the safety of autonomous systems (AS). It is intended to cover the spectrum from those that are semi-autonomous (remotely controlled or operated) through to those that operate largely free of human oversight. Often AS use artificial intelligence (AI), including machine learning (ML), as a key implementation mechanism. The paper relates to the System Assurance and Certification and the AI Safety Foundations focus areas of the AI Safety Landscape. It proposes a novel safety assurance framework.

1.1 Spectrum of Autonomy

In the authors' view AS are best thought of as systems where decisions that would otherwise have been taken by humans are allocated to machines. In practice, many systems have 'shared autonomy' where operators can monitor and potentially over-ride the system's decisions and actions. Often the function (decision) allocation between humans and AS can vary over time. In some sectors, e.g. automotive, the spectrum is codified into 'levels' of autonomy. However, diverse sectors use different definitions of levels, so we choose here just to refer to a spectrum; ultimately our proposed framework is intended to address the whole spectrum of autonomy.

1.2 Scope and Ambition

Producing a complete framework for assuring AS, or even the AI elements of such systems, is a major endeavour. The scope in this paper is more limited and we address:

- Some of the challenges of use of AI and ML in AS, particularly focusing on training and testing;
- Some of the challenges of analysing shared autonomy.

These issues underpin the proposed framework; they are illustrated in section 2 which considers some AS accidents.

AS used in safety-related situations should be subject to safety processes. Previous analysis of quantitative risk analysis (QRA) [Rae et al 2014] identified limitations of classical safety processes, many of which are relevant to AS. Section 3 draws out key issues, considering the impact on both assurance and regulation. It also reviews the literature on AI safety seeking to show a correlation with the QRA limitations.

Section 4 outlines the top-level of the framework and identifies some of the work that will underpin (flesh out) the framework. The framework seeks to address some of the limitations in safety processes identified in the analysis of QRA. The major novelty of the framework is that it includes the use of ML in safety assessment as well as its role in the AS itself.

As the framework is new, it has not yet been applied to an AS. However, a related framework, which has informed our ideas, has been applied in complex healthcare settings, see Section 5. Using such an example also shows how the framework might be used for AI safety more generally, not just for safety of AS.

Sections 6 discusses related work and future plans; Section 7 presents conclusions.

The paper considers some apparently disparate concepts. The ambition is to draw these concepts together to propose a framework that can address the wide range of factors that influence safety and assurance of AS, not just AI. As the ideas are evolving the paper shows a 'direction of travel', not a finished product; it is hoped that this will help stimulate debate.

2 Motivating Examples

Examples are presented which illustrate the problems of the use of AI/ML, and shared autonomy, respectively. The intent is primarily to identify the issues that need to be addressed in producing a safety assurance framework for AS.

2.1 Autonomous Road Vehicles

There have been several fatal accidents with autonomous road vehicles (AVs). We consider two accidents which can be viewed as illustrating some of the issues with AI and ML (as image analysis necessarily uses ML), although this is not fully covered in the accident reports.

A fatal accident occurred in May 2016 when a Tesla on autopilot impacted a truck (semi-trailer) near Williston Florida. The NTSB's analysis of probable cause refers to the driver of the truck failing to yield, and to the Tesla driver's 'inattention due to over-reliance on vehicle automation' [NTSB 2017]. However, the report also says: 'the Tesla's automated vehicle control system was not designed to, and did not, identify the truck crossing the car's path or recognize the impending crash'; if it had recognised the obstacle in its path the Tesla could have applied emergency braking. This shows the challenge of training and testing image analysis systems for operation in an open environment but doesn't give any specific information on the issues for the ML processes.

The NTSB preliminary report on the Uber accident in Tempe Arizona in March 2018 [NTSB 2018] gives a more detailed account of the events leading to the accident than is present in the Tesla analysis. The pedestrian was first observed about 6 seconds prior to the accident. The report says: 'the self-driving system software classified the pedestrian as an unknown object, as a vehicle, and then as a bicycle with varying expectations of future travel path. At 1.3 seconds before impact, the self-driving system determined that an emergency braking maneuver [sic] was needed'. The Uber system relied on the driver for emergency braking and was not intended to initiate an emergency stop.

This example illustrates more fully the problems for ML: how to reliably detect and classify objects in the environment whilst avoiding 'false alerts' that would prevent expeditious progress? It also illustrates the shared autonomy problem, but this is shown more fully below.

2.2 Unmanned Air Systems

The United Kingdom (UK) Army uses an Unmanned Air System (UAS) known as Watchkeeper (WK) for reconnaissance. WK is operated and supported by a ground crew at a Ground Control Station (GCS) but some functions are autonomous, e.g. the ability to execute a go-around (rather than landing). WK does not use AI or ML but it is a good example of the shared autonomy problem.

WK has suffered five accidents to date; a far higher loss rate than the safety analysis predicted [Wilby 2019]. The UK Defence Safety Authority (DSA) report on the fourth WK crash [DSA 2019] draws out three 'themes' from all five accidents (the fifth is still being investigated but the DSA has visibility of the investigation). The themes highlighted are:

1. The incomplete understanding of the full system, and how sub-systems integrate;
2. The need to improve collection and analysis of data;
3. Ground crew and engineer workload.

For example, in the third area, the report cites the high rate of warnings, cautions and advisory (WCA) notifications creating a high workload. Further, the ground crew rely on their collective knowledge to understand the situation and how to respond, e.g. to interpret WCAs rather than referring to documentation (paper or electronic). Thus, based on our reading of the reports and from discussions with the manufacturers [Wilby 2019], we propose the abstract model of accident causation set out in Figure 1.

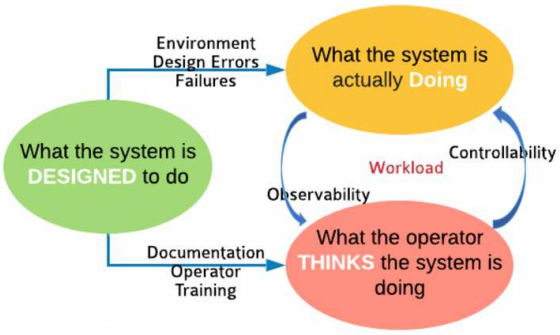

Figure 1. Abstract Model of WK Accident Causation

The design model used, including for safety analysis, was not an accurate predictor of the actual behaviour of the system (in its operational environment). The ground crew training and documentation did not help them to understand the actual system behaviour, including WCAs. Finally, workload had a significant impact on the operators' ability to observe the state of the system and to control it. There is a dissonance between the three sub-models in Figure 1 which is redolent of the distinctions made between 'work as imagined' and 'work as done' [Hollnagel et al 2006]. This dissonance is only likely to get worse for systems that learn in operation.

3 Safety and Assurance Challenges for AS

By assurance we mean justified confidence or certainty in a system's capabilities, including its safety. The motiving examples have given some practical illustrations of why assuring safety of AS is challenging. This section considers these challenges more systematically drawing together three different perspectives. First, it considers 'classical' safety processes, but focusing on their limitations. Second, it considers the 'AI perspective' on safety. Third, it presents a perspective on assurance challenges from a programme addressing assurance and regulation of AS, including their use of AI/ML.

3.1 Safety Processes and their Limitations

The principles of classical safety processes apply to AS. It is necessary to identify hazards, to assess (safety) risks associated with the AS, and to change the design (or operational profile) if the risks are too high, in order to make the system acceptably safe. In this paper we assume that the safety process results in the production of a safety case – a justification (or argument), supported by evidence, that the system is safe to operate in its intended context of use.

A systematic review of Quantitative Risk Analysis (QRA) [Rae et al 2014] showed that the quantitative results are normally not accurate. Safety analysis is not just quantitative, but the problems identified affect the validity of all aspects of safety analysis, whether or not the results are quantified [Rae et al 2014]. The paper proposed a five-level hierarchy for improving QRA, which is also relevant to AS:

1. Unrepeatable, e.g. analysis relies on undocumented assumptions;
2. Invalid, e.g. for WK, design models were not accurate predictors of operational behaviour;

3. Valid but inaccurate, e.g. inadequate treatment of uncertainty such as in object classification;
4. Accurate but challengeable, e.g. insufficient scientific knowledge for parts of the analysis (this is likely to be a particular issue for AS using ML);
5. Ideal (which the paper views as unattainable).

The intention is that flaws at level 1 need to be corrected before a safety analysis can move to level 2, and so on. Flaws identified in [Rae et al 2014] are referred to here as §x.y meaning flaw y at level x, e.g. §2.3 is 'mismatch between the risk assessment and reality' – a key point drawn out in our motivating WK example. The level 4 flaws are of particular relevance to AS – in general we do not have a good scientific basis for safety assessment of ML. [Rae et al 2014] identify other flaws that motivate our framework, and we discuss some of these in Section 4.

3.2 Challenges of AI Safety

The AI community have recently produced a number of papers on the problems of 'AI safety' e.g. [Domingos 2012], [Feige 2019]. It should be noted here that there is a dissonance between the use of the term 'safety' in the AI and safety engineering communities. The latter focus on injury or death to humans; the former takes a wider definition that encompasses robustness, privacy and security, and they tend to focus more on *mechanisms* for producing 'unsafe' results whereas safety engineers are more concerned about *effects*. Fortunately, these different perspectives are complementary and safety engineers can usefully consider 'AI safety' concerns as potential causes of a hazard.

One of the more influential papers [Amodei et al 2016] identifies 'concrete problems in AI'. The paper is largely focused on reinforcement learning; the problems they identify can be rephrased as follows, to make them more general:

1. Avoiding negative side effects – ensuring that the behaviour meets safety constraints;
2. Avoiding reward hacking – ensuring that the behaviour doesn't get a high reward by 'cheating' and thus avoiding the benefit that was sought;
3. Scalable oversight – this concerns the ability of the human to interact with the system both to monitor it, and to respond to requests for confirmation of decisions prior to actions being taken;
4. Safe exploration – if the system can try out new things (new strategies) how can negative outcomes be avoided if these are circumstances that have not been considered hitherto;
5. Robustness to distributional shift – how the system adapts to changes in the operational environment.

Problem 3 arises for WK even though it does not use AI. The two AV accidents illustrate both problems 1 and 3 (although there are issues of psychology here too). The Tempe Uber accident seems to reflect problem 4. There are real-world examples of both problems 2 and 5, e.g., ship classification algorithms that work in the Baltic but not near the equator due to differences in the angle of elevation of the sun, typical wave patterns, etc.

3.3 Assurance and Regulation

The Lloyd's Register Foundation review of robotics and AS (RAS) identified challenges (they used the term 'white spaces') in assurance and regulation of RAS [LRF 2016]. The Assuring Autonomy International Programme that was set up in response to the above review has amplified on these issues and identified Critical Barriers to Assurance and Regulation (CBARs) [AAIP 2018]. CBARs are characterized as problems that must be solved otherwise there is a risk that:

1. Unsafe systems are deployed, or
2. Safe systems cannot be deployed.

The distinction between these two cases rests on the regulatory regime; a permissive regulatory regime which allows systems to be deployed in the absence of contrary evidence is prone to the former risk (arguably this is what happened in the case of the AV accidents outlined above); a restrictive regime is prone to the latter which might inhibit the beneficial use of RAS (hereinafter AS for brevity).

CBARs are intended to highlight key issues in the assurance and regulation of AS, with the aim of focusing research on these issues to expedite safe deployment of AS. The CBARs are intended to apply across different application domains, e.g. UAS or medical devices. Two CBARs relevant to this paper are (simplified and merged from [AAIP 2018]):

1. Monitoring and handover – how can it be ensured and assured that operators retain sufficient levels of attention and concentration to handle problems that arise?
2. Training and Testing – how can it be shown that the training and test sets used in ML give enough coverage of the environment to provide sufficient evidence to allow controlled use of the AS?

All three accidents discussed in Section 2 illustrate the monitoring and handover CBAR; the AV accidents illustrate the second CBAR.

4 The Proposed Framework

The framework proposed here is intended to provide a basis for assurance and regulation of AS, taking into account the use of AI/ML in their development. Although safety principles apply to AS the analysis can be difficult, especially where the behaviour of the system can evolve in operation (e.g. due to ML). The core difficulty is that traditional safety processes assume that safety can be assessed prior to deployment and the assessment remains largely valid through life. The framework therefore includes continued and proactive assessment in operation – in contrast to current safety management that tends only to update safety assessments in response to problems or accidents. It draws on Hollnagel's 'work as imagined' and work as done', but re-setting these as the 'world as imagined' (as we have to consider the system and its environment) and the 'world as observed' (or the 'data world') reflecting the need to analyse operational data to understand and control the dissonances identified above.

4.1 Framework Overview

The framework has four major elements which are conceptually distinct, although they physically overlap to some extent:

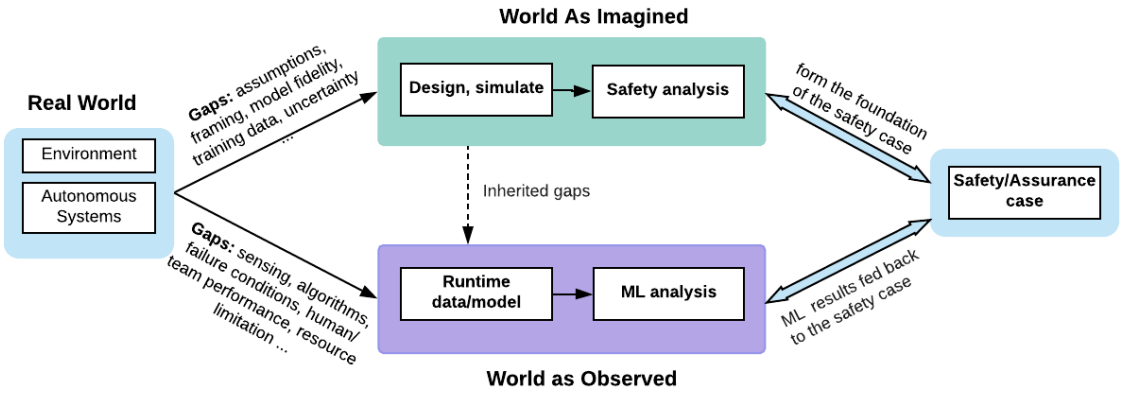

Figure 2. Top-level of the Proposed Safety Assurance Framework

- The 'real world' contains the AS in its operational environment;
- The 'world as imagined' contains design and simulation models, and the results of safety analysis based on those models;
- The 'world as observed' contains the operational data, e.g. images produced by sensors, and the results of ML analysis of the operational data;
- A safety case, initially reflecting just the 'world as imagined' later updated to respond to the 'world as observed', reducing the gaps between the safety case and reality.

The 'real world' environment includes both the physical infrastructure, e.g. roads, and people, including operators, passengers, etc., as well as the AS. This is the 'ground truth'.

Design produces models of the system and environment, including simulations both to define what is to be produced, and to analyse it. In the framework, safety analysis including hazard analysis and the identification of risk mitigations are key elements of the 'world as imagined', but this is where model mismatches (dissonances) can start to arise.

There are 'gaps' between the 'real world' and the 'world as imagined', see Figure 2, that are challenges for safety analysis. The gaps identified in Figure 2 (with example mapping to flaws in [Rae et al 2014]) include:

- Assumptions – what is taken as a 'given', about the system or environment, e.g. driver will be able to intervene and apply emergency braking [§2.3 'mismatch between the risk assessment and reality'];
- Framing (scoping) – deciding what aspects of the 'real world' should be considered, and which do not, e.g. in determining what objects might need to be classified by an AV vision system, such as pedestrians pushing bicycles [§2.2 'Major omissions in the analysis'];
- Model fidelity – precision and accuracy of the models used, including simulations, e.g. point mass approximations to the dynamics of AVs [§2.4a 'Models are used outside their valid scope'];
- Training data – limitations of the data used to train the ML in AS, e.g. biases in data sets used for pedestrian detection, so that some ethnic groups are misclassified [§2.8 'failure to report limitations'];
- Uncertainty – the inability to model all the likely variations in the environment, or the distribution of possibilities, due to 'real world' complexity, [§3.3 'insufficient characterization of uncertainty'].

These factors (and more) contribute to the mismatch between the real and imagined worlds, and thus limit the fidelity (predictive accuracy) of the safety analyses. The training data 'gap' clearly links to the 'training and testing' CBAR.

The 'world as observed' is also affected by 'gaps' as shown in Figure 2, which represent different sorts of mismatch with the 'real world'. These include:

- Sensing – sensor sets, e.g. lidars and cameras, will have limited capabilities and are affected by the environment, e.g. rain, dust clouds, etc.;
- Algorithmic limitations – characteristics of algorithms, e.g. balancing false positives against false negatives, or sensor data fusion not always selecting the best combination of data sources to use;
- Failure conditions – hardware can fail, and this can affect behaviour (but it should be addressed by conventional safety processes);
- Human/team performance – limitations in human cognition and capability, allowing for factors that shape human performance, e.g. the high level of WCAs that affected the WK accidents;
- Resource limitations – many AI/ML algorithms are non-deterministic and, for example, it is possible for object classification algorithms to 'lose' objects if they run out of time between video frames (this is viewed as distinct from algorithmic limitations if the algorithm would not 'lose' the objects given sufficient processing resources).

The human/team performance 'gap' relates to the monitoring and handover CBAR.

The sensor and algorithmic issues are not always disjoint and, for example, the failure of the Tesla to identify the semi-trailer in the Florida accident, is likely to have been due to a combination of such limitations.

As with the mismatches between the 'real world' and 'world as imagined', these gaps are illustrative, not exhaustive. Again, these reflect problems identified by [Rae et al 2014], viz: Sensing maps to §3.1 'insufficient rigour in selecting source data' and Algorithmic limitations maps to §3.2

'incorrect processing of data', albeit extending these categories to the operational system, not just the safety analysis. Note that gaps can also arise in the 'world as observed' from the 'work as imagined' shown as 'inherited gaps' in Figure 2.

The most novel part of the framework is the ML analysis in the 'world as observed'. Safety analysis on the 'world as imagined' is hypothetical – it is based on design models and on an imperfect understanding of the environment. The WK example shows clearly that such mismatches are a serious problem and form significant contributions to the accidents.

Systems employing AI and ML are data rich, and this opens up the possibility of using ML on the operational data both to understand the factors that actually influence safe behaviour, and to validate or to inform refinement of, the safety analyses. The rate at which operational data is produced will exceed human capability for analysis in real-time, but ML offers the opportunity to identify the critical information in the data. In this way the initial safety case would reflect analysis of the 'world as imagined', but then would be updated with ML analysis of the 'world as observed'. Such an approach would allow the safety case to stay in (closer) alignment with the system behaviour, which has been referred to as a dynamic safety case [Denney et al 2015], [Calinescu et al 2018]. The framework also supports feedback, enabling the safety analysis to be improved based on the 'world as observed', and to improve the data collection and ML analysis.

4.2 Assurance of ML in the Framework

The safety case needs to address explicitly each of the 'gaps' identified in Figure 2 and demonstrate that the impact on the behaviour (safety) of the AS is small enough to permit initial operations (see the discussion of regulatory issues below). Figure 2 is not rich enough to show all the issues relating to ML in AS; these will require a more detailed model. The Assuring Autonomy International Programme is working on such models [Ashmore et al 2019]. For simplicity we draw out only a small number of issues highlighted in their ML process model relevant to the 'training and testing' CBAR.

The safety case needs to provide arguments and evidence for the coverage of the training and testing data, noting that the coverage criteria should be informed by the safety process. Thus, for an AV, the focus should be on coverage of those driving situations which are hazardous, e.g. junctions, driving into low sun, and not 'undemanding' situations such as quiet dual carriageways (divided highways). However, it is unlikely (undesirable) that good coverage of near-accident data will be achieved using real world data – it is too rare, and too dangerous to collect – thus collection must be augmented by simulation data to get good coverage of the operational design domain (ODD), from a safety perspective. Thus, the argument should address a number of issues:

- The coverage of the ODD in the data used for training and testing, justifying the 'skew' in the data to get coverage of 'demanding' situations based on the assessment from the safety analysis, e.g. vehicle cut-in and emergency braking to avoid pedestrians;
- The choice between real-world and simulation data, again based on risk assessment, and dealing with the potential problem of distributional shift between the simulator and AV;
- The rationale for choosing factors that shape the learning process including selection of training and test data sets, and 'hyperparameters'.

The 'hyperparameters' in model learning are important as the models learnt are dependent on these values. For example, Bayesian Network (BN) structure learning can be used to identify correlations between elements in the training data. The relevant hyperparameter is Equivalent Sample Size (ESS) which is used to guide the learning process when BDeu (Bayesian Dirichlet equivalent uniform) score is used for the structure learning. In general, increasing ESS leads to learning more links in the structure, but the number of links does not necessarily increase monotonically. The arguments will be domain specific, e.g. for a data set with a skewed distribution, which is very likely with AVs, for example, a smaller ESS value is preferred [Steck 2012].

The Assuring Autonomy International Programme is developing a Body of Knowledge (BoK) which includes template assurance arguments to address these issues [AAIP 2019]. The BoK contains guidance on the structure of assurance arguments and the criteria for evidence and will evolve to provide more details. e.g. using different search algorithms for learning BN structures to improve confidence.

4.3 Example using ML for Safety Analysis

The development of the framework presented here has been influenced by parallel work on safety of medication management which has had to address very similar issues of mismatch between the 'world as imagined' and 'world as observed'. The following illustrative example focuses on the 'ML analysis' in the 'world as observed, as this is the most novel part of the framework.

The example is of a Health IT (HIT) system rather than an AS [Jia 2019]; as mentioned above, this example is chosen as it best illustrates the framework, in the absence of an application to AS. In the HIT work, the mapping to the framework is as follows:

- Real world – hospital environment, including HIT systems;
- World as imagined – healthcare pathway model related to post-operative care of patients following oesophagectomy, and safety analysis using Software Hazard Analysis and Resolution in Design (SHARD) [Pumfrey 1999] based on the healthcare pathway;
- World as observed – real data from the HIT system, and BN structure learning to explore the correlations between the different factors identified in the safety analysis, representing hazards, causes and their effects; this enables the safety analysis to be validated in the real world.

The structure learnt showed new patterns in the ways that nurses carry out their work during post-operative care following oesophagectomy which was not expected as it was neither apparent in the pathway model nor in the safety analysis. This can be characterized as 'flaw' §2.3 [Rae et al 2014] 'mismatch between the risk assessment and reality'.

BN parameter learning was used to quantify the effects of different causes and controls on the associated hazard in order to understand their significance, enabling the introduction of the potential hazard controls to be prioritized.

4.4 Regulation based on the Framework

As noted above, regulatory processes assume that system approval is essentially static, only updated for major events, e.g. significant design changes or accidents. Where AS evolve their capability over time, e.g. using ML, then a more evolutionary approach is needed, with two primary phases.

First, initial operation of the AS should be based on safety analysis of the 'world as imagined'. The decision to approve the system must be based on an assessment of risk that will reflect the arguments that the 'gaps' identified in Figure 2 have been adequately controlled. For AVs this amounts to deciding whether or not the use of the AV in an initial ODD carries an acceptable risk. This will, for example, include consideration of sufficiency of the training and augmentation data in the specified ODD, as outlined above.

Second, operation of the system will provide data, which is then analysed using ML, to either confirm the safety analysis in the 'world as imagined' or to identify areas in which the safety analysis is not consistent with the 'world as observed' (or the 'real world') in a way that is safety significant. The operational data would enable identification of structural weaknesses or 'flaws' in the safety analysis, such as identified in [Rae et al 2014]. For example, if using BN structure learning, the learnt structures may show correlations of causal factors of hazards not reflected in the safety analysis, which is an example of §2.2 'Major omissions in the analysis'. Identifying such a problem should prompt review and, if appropriate, revision to the system design, its operation, e.g. limitations to ODDs for AVs, and update of the safety analysis to reflect a better understanding of the 'real world', etc.

In the WK example this might involve an analysis of ground crew workload (using ML) and a redesign of WK (or more likely its support systems, e.g. the GCS) to reduce workload, with an associated update to the safety analysis and the safety case. Put another way, the ML analysis in the 'world as observed' enables organisations to learn from experience, and to update their approach to safety, rather than the safety analysis being 'open loop' [McDermid 2017].

Of course, the analysis of the 'world as observed' may show that the risk is lower than predicted and that there is a 'safety margin'. In principle, this can be used to justify extending the use of the AS – perhaps expanding the ODD for an AV, or expanding the fleet size for a UAS, or allowing UAS to fly in more congested airspace.

ML can be applied to data from the AS in (near) real-time so, in principle, the safety case can be updated continuously. However, unless regulatory approval can be automated so it can track the evolution of the evidence base, as might be done using run-time certification [Rushby 2008], it is likely that human regulators would approve changes to usage in 'increments' from time to time. This could be done where the evidence produced by ML analysis in the 'world as observed' enables the gaps affecting confidence in the AS to be reduced. In some ways, this is what AS developers are doing now, e.g. progressively expanding the ODDs of AVs. Doing this 'formally' through a regulatory process would be a substantial culture shift, in many domains, not least for the regulators.

5 Discussion and Future Work

The ideas presented here are evolving and have their roots in the earlier analysis of QRA and considerations of how safety engineering can 'catch up' with design engineering [McDermid 2017]. They were crystalized by discussions of the WK accidents. The aspiration is to use ML to understand and reduce the 'gaps' identified in Figure 2. Our initial work on HIT is encouraging, but the ideas need to be applied to AS. There is work using ML on operational data, e.g. for assessing drowsiness of car drivers [Schwarz et al 2019]. However, we believe that use of ML on operational data to update safety cases is a unique perspective, but one that we hope might help to build bridges between the safety and AI communities.

Also, there is a need to revise safety analysis processes to reflect the characteristics of AS and ML, for example changing the notion of controllability used in risk assessment for road vehicles as they become more autonomous [Habli et al 2016]. There is other work in this area, and several authors have made use of Leveson's STAMP/STPA or referred to its utility [NHTSA 2018], [NASA 2018]. STAMP/ STPA reflects a control systems perspective which is highly appropriate for AS, and some aspects of the models might help address the 'flaws' in QRA [Rae et al 2014], however it is not clear that the approach helps particularly with the 'gaps' identified in Figure 2. Thus, we believe that more will need to be done to produce an effective safety analysis process for the 'world as imagined', and thus improve the safety analysis of AS. There is relevant work for AVs, including on Safety of the Intended Function and on a 'standard for safety for the evaluation of autonomous products' by the Underwriters Laboratory (UL), known as UL 4600 [Koopman 2019]. UL 4600 will explicitly address some of the requirements on safety cases for AVs employing AI and ML discussed above; a public draft of UL 4600 should be available during 2019.

We think that the notion of 'gaps' is helpful in considering the wider issues of AS safety. Moving decision-making to an AS can create a 'responsibility gap' where it is unclear who is responsible for an action (in an ethical sense) [Porter et al 2019], and there may be no-one responsible in a legal sense (e.g. Arizona determined that Uber did not have a case to answer for the Tempe fatality). We see the possibility of developing a new and rich socio-technical 'theory of safety' that draws together multiple disciplines, including AI, system safety engineering, law and ethics. A key part of this will be reconciling system safety with AI safety – combining system safety's view of the harm to humans from AS with the understanding from the AI safety community of what might go wrong in AI and ML.

6 Conclusions

The introduction of AS has the possibility of providing significant benefits to society, for example in social care and in

reducing fatalities on the road. However, there are also fundamental challenges to safety and regulatory processes as the WK example, and accidents and incidents with AVs, show.

The framework presented here reflects both empirical understanding of problems with particular AS, and a much more thorough analysis of weaknesses of safety processes in general, and QRA in particular. However, it is abstract, and more detail is needed; the Assuring Autonomy International Programme BoK [AAIP 2019] and standards such as UL 4600 should provide some of the necessary underpinning.

There has been an initial validation of the framework on a healthcare example. It is hoped that the proposed framework will help to unify perspectives on AI safety and contribute to the development of the AI Safety Landscape.

Acknowledgements

This work was supported by the Assuring Autonomy International Programme.

References

[AAIP 2018] Assuring Autonomy International Programme, Critical Barriers to Assurance and Regulation. https://www.york.ac.uk/assuring-autonomy/projects/barriers-to-assurance/ (accessed May 2019)

[AAIP 2019] Assuring Autonomy International Programme, Body of Knowledge. https://www.york.ac.uk/assuring-autonomy/body-of-knowledge/ (accessed May 2019)

[Amodei et al 2016] Amodei, D., Olah, C., Steinhardt, J., Christiano, P.,, Schulman, P., and Mané, D., Concrete Problems in AI Safety, *arXiv.org,* 2016.

[Ashmore et al 2019] Ashmore R., Calinescu, R., Paterson, C., Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges, *arXiv preprint, arXiv: 1905.04223*.

[Calinescu et al 2018] Calinescu, R., Weyns, D., Gerasimou, S., Iftikhar, M.U., Habli, I. and Kelly, T., 2018. Engineering trustworthy self-adaptive software with dynamic assurance cases. *IEEE Transactions on Software Engineering,* 44(11), pp.1039-1069.

[Denney et al 2015] Denney, E., Pai, G. and Habli, I., 2015, May. Dynamic safety cases for through-life safety assurance. *In 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering* (Vol. 2, pp. 587-590). IEEE.

[Domingos 2012] Domingos, P., 2012. A Few Useful Things to Know about Machine Learning, Comm. ACM, Vol. 5, Iss. 10, pp 78-87, October 2012.

[DSA 2019] Defence Safety Authority, Service Inquiry, Loss of Watchkeeper (WK043) Unmanned Air Vehicle over Cardigan Bay in West Wales 24 Mar 17, DSA/DAIB/17/006, 2019.

[Feige 2019] Feige, I., What is AI safety? https://faculty.ai/blog/what-is-ai-safety/ (accessed May 2019)

[Habli et al 2016] Monkhouse, H., Habli, I. and McDermid, J., 2015, September. The Notion of Controllability in an autonmous vehicle context. In *CARS 2015-Critical Automotive applications: Robustness & Safety*.

[Hollnagel et al 2006] Hollnagel, E., Woods, D.D. and Leveson, N., 2006. *Resilience engineering: Concepts and precepts.* Ashgate Publishing, Ltd.

[Jia 2019] Jia, Y., 2019, Improving medication safety using machine learning. *AIME 2019*, Poland, Springer, In press.

[Koopman 2019] Koopman, P., 2019. Private communication.

[LRF 2016] https://www.lrfoundation.org.uk/en/news/foresight-review-of-robotics-and-autonomous-systems/ (accessed May 2019)

[McDermid 2017] McDermid, J., Playing Catch–Up: The Fate of Safety Engineering? *Developments in Systems Safety Engineering*, *Safety-Critical Systems Club*, M Parson T P Kelly (Eds), 2017.

[NASA 2018] Alves, E.E., Bhatt, D., Hall, B., Driscoll, K., Murugesan, A. and Rushby, J., 2018. Considerations in Assuring Safety of Increasingly Autonomous Systems.

[NHTSA 2018] National Highway Traffic Safety Administration, Functional Safety of an Automated Lane Centering System, DOT HS 812 573, 2018.

[NTSB 2017] National Transportation Safety Board, 2017. Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida May 7, 2016, NTSB/HAR-17/02

[NTSB 2018] National Transportation Safety Board, 2018, Preliminary Highway Report, HWY18MH010.

[Porter et al 2018] Porter, Z., Habli, I., Monkhouse, H., & Bragg, J., 2018. The moral responsibility gap and the increasing autonomy of systems. In International Conference on Computer Safety, Reliability, and Security (pp. 487-493). Springer.

[Pumfrey 1999] Pumfrey, D.J., 1999. The principled design of computer system safety analyses (Doctoral dissertation, University of York).

[Rae et al 2014] Rae, A., Alexander, R. and McDermid, J., 2014. Fixing the cracks in the crystal ball: a maturity model for quantitative risk assessment. *Reliability Engineering & System Safety*, 125, pp.67-81.

[Rushby 2008] Rushby, J., 2008, March. Runtime certification. *In International Workshop on Runtime Verification* (pp. 21-35). Springer, Berlin.

[Schwarz et al 2019] Schwarz, C., Gaspar, J., Miller, T., Yousefian, R., 2019. The Detection of Drowsiness Using a Driver Monitoring System, ESV 2019.

[Steck 2012] Steck, H., 2012. Learning the Bayesian network structure: Dirichlet prior versus data. *arXiv preprint arXiv:1206.3287.*

[Wilby 2019] Wilby, A., 2019. Private communication.