The State of Solutions for Autonomous Systems Safety

Rob Alexander[1], **Rob Ashmore**[2], **Andrew Banks**[3]
on behalf of the SCSC Safety of Autonomous Systems Working Group (SASWG)

[1]University of York, UK

[2]Dstl, UK

[3]Frazer-Nash Research Limited, UK

Abstract *Autonomous Systems are seeing increasing use and increasingly safety-significant application. Consequently, the safety of autonomous systems is an important topic. To reflect this importance the Safety Critical Systems Club (SCSC) has established the Safety of Autonomous Systems Working Group (SASWG). This paper introduces the SASWG and describes (and justifies) the approach it is taking. A running example is used to illustrate challenges, which are organised against three "difficulty horizons". Potential solutions to some of the challenges are outlined; possible research directions are suggested for other challenges. Some proposed but invalid solutions are also identified. Overall, whilst the SASWG acknowledges the very significant benefits that could accrue from autonomous systems, it believes their development and implementation should be pursued carefully and thoughtfully.*

1 Introduction

The SCSC has created a working group to help with the autonomous system safety problem. Autonomous systems (AS) safety is important. There is great deal of money and effort going into developing, manufacturing and field-testing autonomous systems of many kinds. Uber, Waymo, Tesla and others are experimenting with autonomous cars. DeepMind is developing autonomous diagnostic

systems for use in hospitals. This volume and intensity of research and development is very different to ten or even five years ago.

In response to this, the Safety Critical Systems Club (SCSC) has established the Safety of Autonomous Systems Working Group (SASWG). Our long-term objective is to produce clear guidance on how AS and autonomy technologies should be managed in a safety-related context, throughout the system lifecycle, in a way that is tightly focussed on challenges unique to autonomy.

The creation of SASWG is timely because, as noted above, such systems are starting to be used, are starting to be involved in accidents, and are starting to attract specific public controversy. The Tesla crash of May 2016 (National Transportation Safety Board 2017) has shown the public one of the risks in shared-autonomy control. DeepMind's work with patient data at the Royal Free Hospital has attracted controversy (Mathieson 2017), as have proposals for platooning of heavy goods vehicles on motorways (Department for Transport 2017).

The SASWG is also timely because some AS technology developers, system integrators and manufacturers are now taking AS safety seriously. Many integrators and manufacturers are working with regulators and legislators to find ways for their products to be approved; many technology developers are working hard to convince integrators that their components can be assured safe. For example, NVIDIA claim that the "Parker" system-on-a-chip has features that support claims of compliance with ISO 26262 (Shapiro 2016).

The picture is not always so clear. It is difficult to determine what safety processes were associated with the decision to use DeepMind's human health prognostic software. We can also observe that for certain key technologies, such as neural networks, there are no extant high-integrity libraries available.

In this paper, we introduce the SASWG in terms of our proposed general approach and our arguments for its appropriateness. We then describe some relatively easy, moderately difficult, and possibly unsolvable challenges, and say how we intend to act on each of them.

To some extent, this paper provides a vision of how the SASWG will operate over the next few years. However, in doing so it offers an implicit answer to the broader question of how we can advance system safety for radically novel technology that seems to be rapidly being deployed in society.

2 Progress

We have made some progress in structuring the problem. The SASWG's long-term objective is to produce clear guidance on how AS and autonomy technologies should be managed in a safety-related context, throughout the lifecycle, in a way that is tightly focussed on challenges unique to autonomy. This objective includes, but is not limited to, guidance on how to achieve:

- Safety management;
- Safety assurance;
- Communication of safety (and safety limitations) to end users.

Communication with a range of stakeholders is obviously important. In the above list we have chosen to specifically highlight end users as they may have an important role in ensuring overall system safety, for example, by monitoring the behaviour of an autonomous component.

We have adopted a set of principles to guide our work. These do not fully describe what we do: rather, they guide us to avoid well-known traps we could fall into. In the SASWG we have experience with initiatives that failed, often drowning in a mire of complexity or vague generality. Each principle is a response to a problem that someone else has seen.

Our principles include:

- **The distinction between autonomous and automatic is not useful.** We are interested in systems where it is currently a challenge to make a convincing safety argument. We will favour working by example over abstract discussion, even if at times those examples are fictitious and overly simple.

- **Consider the whole system lifecycle, but don't try to impose our own lifecycle model.** We recognise that developers use a wide variety of different lifecycles. To the extent that we impose a very specific lifecycle or top-level process, our guidance will be impractical for many of them. We acknowledge, however, that AS software may be updated much more frequently than is currently the case, for example, for aircraft systems.

- **Focus on problems specific to autonomous systems.** Safety-critical AS inherit essentially all the safety engineering concerns of conventional systems. It is easy to get distracted by this, especially when intimidated by the difficulty of the obviously AS-specific challenges. However, if the SASWG spends time thinking about well-established safety practice, and modifying it only slightly for AS, then it is wasting its members' time. If the SASWG publishes guidance which purports to be about AS safety, but spends many words on established system-safety practice, it is wasting the time of others.[1]

[1] For an illustration of this problem, see the criticism in (Alexander, Kelly, and Herbert 2009) of the *Unmanned Systems Safety Guide for DoD Acquisition* (United States Department of

- **Focus on the challenges raised by specific tasks, environments and solutions.** In a novel field, solutions are likely to be most easily found in specific cases. Generalisation of solutions (e.g. to "all autonomous systems" or "all systems of autonomy level X") may be possible later, but is rarely practical until many special-case solutions have been found.

- **Prioritise understanding of requirements over understanding of how we might implement and assure them.** In the first instance, the most important thing to do is to define the general requirements that an AS must meet in order to be safe. In other words, we will adopt the DO-178C and ISO 15288 / ISO 12207 model of providing objectives rather than means of compliance. In our process of working out those requirements, however, we will consider concrete implementations and technologies (as we do in later sections of this paper). Inevitably, given the maturity of the field, potential means of compliance may be immature and unproven. Where what we are asking is currently impossible, we will make it clear that we know this (e.g., by placing such requirements in a separate section of a document).

- **Don't duplicate existing work.** Wherever possible, reference out to existing and emerging standards, guides and other sources, rather than incorporating their detail in our guidance. If we establish that our guidance, applied to examples, generates requirements that can be discharged by following existing guidance and standards (e.g. software safety requirements that look to be dischargeable under DO-178C or ISO 26262 / IEC 61508), then we should give no further guidance. Where emerging standards are used, care will be taken to ensure they are appropriately validated.

- **Consider security, but only insofar as it affects safety.** We acknowledge that a system can only be safe if it is also secure (and that a safe system can be made unsafe by the addition of security features). However, AS raise many novel security challenges, including ones relating to data rather than software. These specific challenges will be work enough for us; covering the entire security space within the SASWG is not feasible.

Defense 2007). Note, however, that the former assumes the latter was written for a target audience of safety engineers, which is probably not correct.

- **Consider ethics, but only in so far as it is relevant to safety, and without commitment to a particular ethical system or rules.** Ethical decisions are important, but are outside the expertise of the SASWG. Instead, our aim is to understand the types of safety requirement that ethical considerations may introduce.

3 Challenges

We can identify challenges of different kinds. Different kinds of challenges require different solutions. There are many ways to subdivide the space of "AS safety challenges", but we will offer just one here —challenges of different apparent difficulties. We refer to these as "horizons" (which are orthogonal to the "development and operation" subdivision):

- Problems that we are quite close to solving;
- Problems where it is not clear how to solve them;
- Problems where is it not clear that they can be solved at all.

To help illustrate these problems over subsequent sections, we develop a fictitious example. Specifically, our running example is a medical imaging system that makes treatment decisions based on a learned model.

Figure 1 below is a schematic of the process supported by the AS. In outline: a patient's records declare that an image-based diagnosis is required; the patient is scanned (via a non-invasive procedure); the resulting image is fed to an automated diagnosis system, a component of which is the AS; the resulting diagnosis is recorded in the patient records.

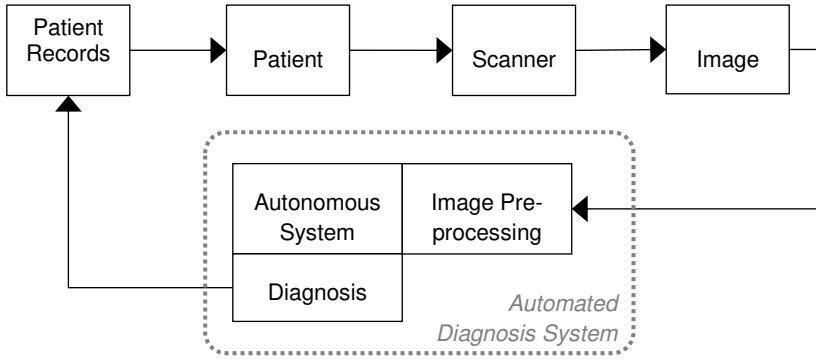

Fig. 1. Schematic of the process supported by the AS

For the purposes of this example, it is assumed that the diagnosis will be one of the following three categories:

- Re-Image, which requests another image be captured (for example, because the supplied image is not of a high enough quality);
- Treatment, which initiates a course of treatment with unpleasant and potentially significant side effects;
- All-Clear, which suggests no further action need be taken.

We make the following assumptions:

- Some pre-processing is likely to be required before the image is fed into the decision-making part of the AS. This pre-processing, which could be implemented using traditional software techniques, is not a focus for this analysis. It is, however, noted that this step offers a means by which the behaviour of the AS could be subverted (Stevens et al. 2017).
- The task being performed by the AS is, in essence, classification based on image recognition. Hence, it is assumed that the AS is implemented using some form of Artificial Neural Network (ANN). That said, many of the assurance challenges would apply regardless of the implementation technology.
- It is assumed the ANN is developed via some form of training, test and validation activity, after which it is left fixed. In particular, when in operational use, providing the same image to the AS will always result in the same diagnosis[2]. (This assumption does not prevent observations from in-use behaviour being used to develop an improved AS.)

The key sub-property[3] of "suitably safe" relates to whether the correct diagnosis has been provided. Note that there is no always-safe diagnosis — a patient is only optimally safe when a correct diagnosis is made. For example, the AS diagnosing Re-Image rather than All-Clear wastes resources and causes unnecessary worry for the patient. Alternatively, diagnosing Re-Image rather than Treatment

[2] Strictly speaking, it is the automated diagnosis system that produces the diagnosis. However, since the AS is the focus of this activity, within this paper the term AS is used as a shorthand for referring to the wider diagnosis system.

[3] Other important properties include: always making a diagnosis; making a diagnosis within a specific time; etc.

potentially delays treatment, which could have significant adverse effects. More generally, all incorrect diagnoses are associated with some form of harm.

The following sections discuss problems that are faced by safety engineering for the above example. We should note that this is a relatively simple system with a very narrow space of perception and action; unrealistically so, at least given current and likely-near-future imaging technology. As such, however, it presents an intriguing challenge: if we cannot demonstrably make this safe, then we may have little chance of making autonomous cars (or autonomous robot surgeons) safe at all.

3.1 Close to solving

There are problems that we are quite close to solving. Problems in this category are beyond current practice, but there seems to be a plausible path to their solution. Typically, we already have the frameworks and techniques we need; we get general agreement about, or empirically determine, some aspects, and then apply them.

3.1.1 Running example

Our main example here is checking that the inputs seen during operational use are appropriately similar to the distribution of Training, Test and Validation (TTV) data. The "appropriately similar" here, formally, is that the operational data (both individual images and collections of images) should be statistically indistinguishable from the TTV data.

The medical diagnosis device is of course part of a controlled process (e.g. only images from defined-suitable medical scanners need analysing), which means that the operational environment of this AS is much simpler that associated with, for example, an autonomous car.

Put simply, the challenge here is just about comparing two statistical distributions. There are at least three distinct types of distribution shift that can occur between TTV data and operational data (Moreno-Torres et al. 2012). In simple terms, for classification problems[4], they are:

- Covariate shift relates to changes in the independent (or input) variables;

[4] That is, problems where the aim is to allocate an input to a given class, for example, to take pictures as input and identify all "pictures of cats".

- Prior probability shift relates to changes in the target (or output) variable;
- Concept shift relates to changes in the relationship between the independent and target variables.

Techniques exit to detect each of these, but how to do them adequately in a safety-critical context needs to be determined.

3.1.2 General discussion

For this problem, the SASWG approach is:

1. Define the need for this as precisely as possible (so that solutions can be proposed and evaluated) i.e. saying how requirements can be stated;
2. Identify possible published solutions and directing attention of developers to those;
3. Note that these topics are probably best solved by developers attempting to do so as part of new product development, and regulators attempting to negotiate how they will accept them.

There are many similar problems. For example:

- Putting in a human-level monitor that checks one in every x diagnoses (for a suitable value of x) is easy enough. In many ways this could be viewed as being similar to normal quality control sampling, for example, in a biscuit factory. This could provide quantitative evidence as to how often the human and the AS disagree. This, however, implicitly assumes the human is always correct, which may be a questionable assumption. Longer-term studies, with associated ethical implications, would be needed to properly monitor the AS performance; for example, recording how many patients who had been given the all clear subsequently reported with symptoms. Again, this type of study is similar to those that are, presumably, conducted on a routine basis within the medical world.

- Ensuring that “advisory systems” are indeed merely advisory, and remain so. As we see humans and AS work more closely together, it is possible that safety arguments will make claims that the AS is only advisory. This is an easy claim to make, but it is not always easy to justify. For example, in order for an AS to be advisory: first, the person being advised has to have some independent way of checking the advice / making the decision; second, the “incident investigation” process

has to be able to cover cases where the person didn't follow the advice, bad things happen, yet the person is not automatically castigated for not following the system's advice. Conversely, over time humans may assign too much trust to the AS and, for example, stop cross-checking it.

We probably cannot safely deploy anything of note by only solving the problems in this category. As a society we will need to solve problems in the next category.

3.2 Not clear how to solve

There are many problems that it is not clear how to solve. Problems in this category are beyond current practice (as with the previous category), and here there are many competing approaches with none of them clearly the way forward. Research is needed.

One example is establishing adequate standards of Verification and Validation (V&V) for safety-related use of offline learning. In order to have confidence in the predicted performance of a learning system, confidence is needed in the data (noting that confidence in training and test data supports different types of argument than confidence in the validation data).

3.2.1 Running example

For the medical device example, we need:

1. Independence of testers and of the Training, Test and Validation (TTV) data;
2. Confidence in TTV data;
3. A way of ensuring that typical errors have been avoided during the development, implementation and use of an AS.

Independence of testers and test data is superficially straightforward. Once the development team has produced a candidate ANN, this should be subject to independent validation. This involves a separate team analysing the performance of the ANN using data that is separate from that used by the development team for training and testing the ANN. In other words, the training and test data (used for development) and the validation data need to be independent. However, it is not clear what constitutes adequate independence in this context.

For confidence in TTV data, we could rely on historical data that includes both the original image-based diagnosis and the final outcome; these could differ if, for example, treatment was subsequently required despite an initial All-Clear

diagnosis. It would need to contain difficult cases that result from failures in other parts of the process (e.g. images corrupted either during capture by the scanner or in transmission to the ANN); determining the correct diagnoses for these cases could be challenging.

Avoidance of typical errors is analogous to the use of robustness testing in traditional safety-critical software, where attempts are made to induce known problems (e.g. arithmetic overflow). Some possible errors associated with some forms of machine learning are known, for example, their vulnerability to adversarial examples (where a small, carefully chosen, change in input leads to an undesirable change in output (Szegedy et al. 2014), (Goodfellow, Shlens, and Szegedy 2015)). Some level of protection against this phenomenon may be provided by including such examples in the training set and using a larger model. However, what is missing is a generally accepted list of typical errors (possibly organised by machine learning approach), ways of detecting them and ways of correcting them. Even limiting ourselves to modern ANNs, no such list is known to the SASWG. (Amodei et al. 2016) catalogues a wide variety of failure modes for learners in general, but it does not claim to be exhaustive.

The medical device is of course a system where the worst-case single failure is a single unnecessary death. For vehicles, especially ones with many passengers, we will need much higher confidence in their V&V. Statistical techniques are very unlikely to give us adequate confidence here, due to the huge volume of data required. We will need analytical techniques, such as those proposed for (specific subclasses of) ANNs in (Kurd 2005) and (Katz et al. 2017).

General discussion

For this problem, the SASWG approach is:

1. Define the need for this as precisely as possible (so that solutions can be proposed and evaluated) i.e. say how requirements can be stated (recognising, though, that this itself is likely to be very difficult). Where we cannot say how to define the requirements, we will identify the barrier to doing so such that further research on that is possible.
2. Catalogue materials that point towards a solution, where available.
3. Note that applied research (of the type that e.g. Innovate UK would fund) is needed on these topics.

There are many similar problems. For example:

1. The lack of high-integrity frameworks and libraries that implement neural networks. One of the reasons rapid progress is being made in the field of AS is the ready availability of supporting libraries and frameworks (eg, TensorFlow (https://www.tensorflow.org/) and Caffe

(http://caffe.berkeleyvision.org/)). However, these libraries have not been developed using the rigorous processes one would associate with safety-critical software. Furthermore, the role of these frameworks in typical system architectures means it is difficult, if not impossible, to sandbox them and treat them as Software Of Uncertain Provenance (SOUP). It should be noted that this is, in effect, a marketplace problem rather than a technical one. Perhaps in the future we will see "certified" machine learning frameworks, in the same way there are "certified" operating systems today.

2. Assuring driver readiness in SAE level 3 and 4 systems. A feature of these systems is the need for control to revert from the AS to the human driver. Maintaining system safety through this handover and, in particular, ensuring the human is ready and able to take control, and understands that they have done so, is a challenge. This becomes significantly more challenging if the handover has to occur in a critical, or emergency, situation. This is so much of a concern that a recent report from the German Federal Government (The Ethics Commission on Automated and Connected Driving 2017) states: "*The software and technology in highly automated vehicles must be designed such that the need for an abrupt handover of control to the driver ("emergency") is virtually obviated.*"

3. Finding quantitative metrics that can measure the extent to which an AS has been tested and, furthermore, setting appropriate thresholds for different types of AS use. This consideration is, in some ways, analogous to the use of statement, branch and modified condition / decision coverage in DO-178C and ISO 29119. There is some work that suggests, for neural networks, that neuron activation patterns may be a helpful measure (Tian et al. 2017). There are also more general measures that relate to the way that the TTV data covers the input domain, examples include: the gap ratio (Bishnu et al. 2015), which is a summary statistic from the distribution of distances between TTV samples; and the Maximum Empty Hyper Rectangle (MEHR) (Lemley, Jagodzinski, and Andonie 2016) which identifies the largest region of the input domain without a TTV samples. Moving away from machine learning and into a broader space of autonomous applications, we have concepts like situation coverage (Alexander, Hawkins, and Rae 2015) and environmental hazard analysis (Dogramadzi et al. 2014).

A great many AS safety problems are in this category. If we can solve them, we will probably be able to deploy a range of interesting and useful AS in safety-related situations.

3.3 Not demonstrably solvable

Some problems may yet derail the whole endeavour. Problems in this category are not demonstrably solvable, and there are reasons to believe they are not solvable at all.

One example is avoiding pathological learning behaviours in online learning systems. (Amodei et al. 2016) catalogues a range of unsafe failure modes that such systems can exhibit, albeit with a bias towards one specific class of techniques (reinforcement learning) and a very broad definition of "safety". Their failure modes are thus:

- Negative side effects of learned behaviour;
- "Reward hacking", whereby learning systems learn pathological behaviours that nevertheless meet the implemented definitions of good behaviour;
- Systems that require impractical amounts of human oversight;
- Unsafe exploratory behaviours;
- Models become unsafe due distributional shift in inputs or needed outputs.

Some of the above are extensions of challenges we noted earlier. For example, a distributional shift can potentially invalidate the results of any validation, certainly any validation that was efficient enough to be practical.

Some of the problems are exhibited by offline learning systems, but are worse in online ones. For example, negative side effects can be learned offline, but in that case there is potential for them to be detected and fixed offline. Online learning systems can learn to create negative side effects during operation.

For all the above problems, Amodei et al point to work on possible solutions, but few of them are compelling at this time, and none of them are ready for general engineering use.

It may be possible to introduce system-level architectural protection, for example, by implementing a monitor using conventional techniques (Caseley 2016). However, it may be difficult to strike an appropriate balance between allowing an AS sufficient freedom whilst simultaneously implementing a monitor that can provide strong guarantees about overall system behaviour.

For this category of problems, the SASWG approach is:

1. Note their existence, as before;

2. Indicate unequivocally that these are not currently tractable, with clear justification of such (so that developers do not waste effort on impossible things);
3. Provide a high-level, fairly abstract indication of the requirements any solution would need to meet;
4. Note that these problems are of a type that will hinge on basic research, of the kind that UK research councils (such as the EPSRC) fund.

Other problems in this category include:

- Safety arguments are inherently incomplete with respect to "unknown unknowns". Novel threats (and deliberate attacks) can potentially appear at any time, and with novel technology they are more likely than with more established technology. See, for example, the recent reports of tricking autonomous car visual identification by applying small paper stickers to road signs (Snyder 2017). In addition, AS may feature emergent behaviour that was not predicted at design time.
- Interaction between independently-designed AS. Consider a future in which there are a variety of different self-driving cars, which have been produced by competing manufacturers. It may be the case that each car considered in isolation is judged to be (in some sense) "safe" but the combination of different cars leads to unexpected, undesirable and potentially unsafe situations. That is, the system-of-ASs has an unsafe emergent behaviour. Something similar has already been seen, albeit in a non-safety-related context, in the way that simple automatic trading algorithms can lead to flash crashes (Markets Committee, Bank for International Settlements 2017).

  One potential solution could involve a central authority maintaining a detailed simulation of all AS within the system, but even this approach raises many potential difficulties, for example: demonstrating the simulation has sufficient fidelity; choosing an appropriate number of simulation runs, with suitable properties; keeping up-to-date with software (and system) updates; deciding how to resolve identified conflicts (without having an inappropriate benefit for "first to market" manufacturers).
- Providing confidence in new types of hardware. The safety-critical community has traditionally been very wary of new types of hardware, which can induce new types of failure mode. Consider, for example, the challenges associated with introducing multi-core processors (see (FAA

Certification Authorities Software Team 2016) for some details). Yet, multi-core processors are significantly simpler than, for example, neuromorphic chips or system-on-chip solutions like NVIDIA's "Parker".

Although there is not necessarily a link between the level of challenge and the associated reward, if we can solve the problems in this category, we may be able to develop some extremely powerful systems, with concomitant social impacts. We should of course remain aware of the ethics of deploying such systems, as their social impacts may not necessarily be beneficial. Conversely, there may be an ethical imperative to deploy such systems, for example, if they can reduce accidental deaths (The Ethics Commission on Automated and Connected Driving 2017).

3.4 Invalid Solutions

Some proposed solutions are invalid. Beyond the solution ideas identified above, there are many invalid proposed solutions.

Some proposed solutions are plausible in theory but invalid in practice. The most prominent example is the idea that if we get autonomous cars on the roads we can accumulate enough miles for statistical evidence of safety. The problem with this is that such evidence is unstructured; it is evidence when taken as a whole, but we cannot subdivide it, nor relate it to specifics of the system design. This means that if the system is changed in any way then all of the evidence is invalidated. This is, of course, incompatible with any realistic development and maintenance process for a single vehicle model, let alone for the class of autonomous cars in general. There is also the issue that many of the driven miles will be, in essence, "boring". Conversely, we are interested in the behaviour of the autonomous cars in exceptional, or emergency, situations, which by their very nature are likely to be sparsely represented in the sampled driving hours.

Some partial solutions are useful but easy to overhype. For example, being able to explain an individual decision made by an AS is very useful, but it doesn't scale as an option for providing confidence in an AS's overall behaviour. In particular, being able to explain a decision in a human-accessible form (Ribeiro, Singh, and Guestrin 2016) can be useful during system development and accident investigation. However, it will never be possible for a human to inspect enough examples, which cover enough different situations, to provide confidence in overall system behaviour.

Some partial solutions are intuitively compelling, but only because of unsupportable assumptions. For example, it is sometimes suggested that an AS could be assured in the same way that human beings, or working animals such as guide and police dogs, are. Focusing on the human example, their assurance relies on things like training courses, tests and examinations. All of these approaches

implicitly rely on the way that humans interpolate and extrapolate from their training to new and novel situations.

There is evidence, unfortunately, that extant machine learning algorithms interpolate and extrapolate in a different way to humans. We can create adversarial examples (Szegedy et al. 2014), where a small change in input leads to a large change in output (and, crucially, a change that a human would not produce) is well known. As a minimum, this indicates that algorithms would need radically different types of training, tests and examinations than are used for humans; at maximum, it could suggest that this type of approach will not be feasible at all.

The SASWG will systematically catalogue these known invalid solutions, and publish clear proposal-and-rebuttal statements to inform people of the problems with them. Where solutions are partial, we will explain the role that they can legitimately have.

4 Conclusions

Achieving autonomous system safety, particularly in very challenging cases like fully autonomous cars, is likely to be very difficult indeed. However, there are easier cases where some progress appears feasible.

In this paper, we have defined the approach that the SASWG is taking towards tackling the general AS safety problem. Our approach has the potential to help progress. The degree to which we will ever be able to have confidence in the safety of autonomous systems is unknown.

Overall, we concur with the statement made by a member of the board that investigated the Tesla accident in May 2016, and would extend the statement to cover all applications of automation and autonomous systems: "*the potential benefits of automation on our streets and highways are truly phenomenal, but they must be pursued carefully and thoughtfully*" (National Transportation Safety Board 2017).

References

Alexander, Rob, Heather Hawkins, and Drew Rae. 2015. 'Situation Coverage - a Coverage Criterion for Testing Autonomous Robots'. YCS-2015-496. Department of Computer Science, University of York.

Alexander, Rob, Tim Kelly, and Nicola Herbert. 2009. 'A Critique of the "Unmanned Systems Safety Guide for DoD Acquisition"'. In *Proceedings of the 27th International System Safety Conference (ISSC)*.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. 'Concrete Problems in AI Safety'. *ArXiv:1606.06565 [Cs]*, June. http://arxiv.org/abs/1606.06565.

Bishnu, Arijit, Sameer Desai, Arijit Ghosh, Mayank Goswami, and Subhabrata Paul. 2015. 'Uniformity of Point Samples in Metric Spaces Using Gap Ratio'. In *Theory and Applications of Models of Computation*, 347–58. Lecture Notes in Computer Science. Springer, Cham. doi:10.1007/978-3-319-17142-5_30.

Caseley, P. R. 2016. 'Claims and Architectures to Rationate on Automatic and Autonomous Functions'. In *11th International Conference on System Safety and Cyber-Security (SSCS 2016)*, 1–6. doi:10.1049/cp.2016.0855.

Department for Transport. 2017. 'Green Light for Lorry "Platooning" - GOV.UK'. August 29. https://www.gov.uk/government/news/green-light-for-lorry-platooning.

Dogramadzi, Sanja, Maria Elena Giannaccini, Christopher Harper, Mohammad Sobhani, Roger Woodman, and Jiyeon Choung. 2014. 'Environmental Hazard Analysis - a Variant of Preliminary Hazard Analysis for Autonomous Mobile Robots'. *Journal of Intelligent & Robotic Systems* 76 (1): 73–117. doi:10.1007/s10846-013-0020-7.

FAA Certification Authorities Software Team. 2016. 'Multi-Core Processors'. CAST-32A. Federal Aviation Administration.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2015. 'Explaining and Harnessing Adversarial Examples'. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. http://arxiv.org/abs/1412.6572.

Katz, Guy, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. 2017. 'Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks'. In *Computer Aided Verification*, 97–117. Lecture Notes in Computer Science. Springer, Cham. doi:10.1007/978-3-319-63387-9_5.

Kurd, Zeshan. 2005. 'Artificial Neural Networks in Safety-Critical Applications'.

Lemley, J., F. Jagodzinski, and R. Andonie. 2016. 'Big Holes in Big Data: A Monte Carlo Algorithm for Detecting Large Hyper-Rectangles in High Dimensional Data'. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, 1:563–71. doi:10.1109/COMPSAC.2016.73.

Markets Committee, Bank for International Settlements. 2017. 'The Sterling "Flash Event" of 7 October 2016'. Bank for International Settlements. http://www.bis.org/publ/mktc09.pdf.

Mathieson, S A. 2017. 'Pssst... Wanna Participate in a Google DeepMind AI Pilot? Be Careful'. *The Register*. August 23. https://www.theregister.co.uk/2017/08/23/nhs_google_deepmind_lessons/.

Moreno-Torres, Jose G., Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. 2012. 'A Unifying View on Dataset Shift in Classification'. *Pattern Recognition* 45 (1): 521–30. doi:10.1016/j.patcog.2011.06.019.

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. National Transportation Safety Board. https://www.ntsb.gov/investigations/AccidentReports/Pages/HAR1702.aspx.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. '"Why Should I Trust You?": Explaining the Predictions of Any Classifier'. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. KDD '16. New York, NY, USA: ACM. doi:10.1145/2939672.2939778.

Shapiro, Danny. 2016. 'Introducing Parker, NVIDIA's Newest SOC for Autonomous Vehicles | NVIDIA Blog'. *The Official NVIDIA Blog*. August 22. https://blogs.nvidia.com/blog/2016/08/22/parker-for-self-driving-cars/.

Snyder, John Beltz. 2017. 'Researchers Hack a Self-Driving Car by Putting Stickers on Street Signs'. *Autoblog*. August 4. https://www.autoblog.com/2017/08/04/self-driving-car-sign-hack-stickers/.

Stevens, Rock, Octavian Suciu, Andrew Ruef, Sanghyun Hong, Michael Hicks, and Tudor Dumitraş. 2017. 'Summoning Demons: The Pursuit of Exploitable Bugs in Machine Learning'. *ArXiv:1701.04739 [Cs]*, January. http://arxiv.org/abs/1701.04739.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. 'Intriguing Properties of Neural Networks'. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. http://arxiv.org/abs/1312.6199.

The Ethics Commission on Automated and Connected Driving. 2017. 'Ethics Commission — Automated and Connected Driving'. German Federal Ministry of Transport and Digital Infrastructure. Accessed September 22. https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html.

Tian, Yuchi, Kexin Pei, Suman Jana, and Baishakhi Ray. 2017. 'DeepTest: Automated Testing of Deep-Neural-Network-Driven Autonomous Cars'. *ArXiv:1708.08559 [Cs]*, August. http://arxiv.org/abs/1708.08559.

United States Department of Defense. 2007. 'Unmanned Systems Safety Guide for DoD Acquisition'. ADA472102. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA472102