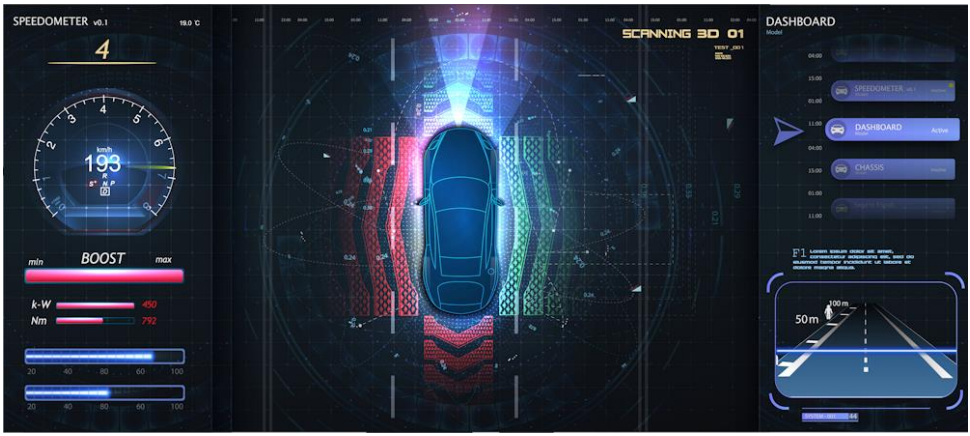
The Safety of Autonomous Systems: An SCSC Working Group



The SCSC Safety of Autonomous Systems Working Group discusses the history and motivation of the group and presents a summary of the significant progress it has made in developing guidance in the uniquely challenging domain of safety assurance of autonomous systems and technologies.

The Group

In January 2016, the Data Safety Initiative Working Group (DSIWG) released version 1.3 of "*Data Safety Guidance*". Motivated by the continued success of the DSIWG, the Safety Critical Systems Club (SCSC) considered the potential for other working groups. Then, as now, rapid developments were being made in Artificial Intelligence (AI) using Machine Learning (ML) techniques. Autonomous systems, enabled by a combination of AI and ML, were becoming more important. Consequently, the Safety of Autonomous Systems Working Group (SASWG) was formed.

The group held its first meeting in January 2017. As is typical for SCSC working groups, this involved a collaboration between industry, academia and government. This balanced input has been critical to the group's success. It combines innovative research with practical concerns.

Following the DSIWG's example, the SASWG coordinates its main outputs with the Safety-critical Systems Symposium (SSS). In January 2018, a paper outlining "*Safety-Related Challenges for Autonomous Systems*", SCSC-143, was released. Developing this paper helped solidify the group's thinking.

Maintaining the annual drumbeat, the next publication, in January 2019 (for release at SSS'19), was "*Safety Assurance Objectives for Autonomous Systems*", SCSC-153 [1]. This established a structure for the group's guidance material. Whilst, in this edition, large portions of the structure were simply placeholder sections, initial guidance relating to the computational aspect of autonomous systems safety was included. Positive feedback was received.

An updated version of "*Safety Assurance Objectives for Autonomous Systems*", SCSC-153A [2], was produced in January 2020 and released at SSS'20. This contains a complete list of 45 objectives, along with guidance on how they may (at least partially) be met. As such, it provides useful guidance for those wanting to develop a safety (or assurance) argument for an autonomous system. Reaching this point just three years after the group's first meeting is a significant achievement.

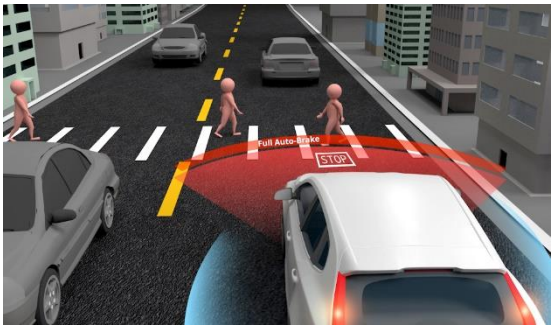
It is a great credit to the SASWG Chair, the volunteers' employers and, most especially, the volunteers themselves that the SASWG has functioned so well and achieved so much.

Like other SCSC working groups, the SASWG is entirely a volunteer effort. The group relies on the generosity of its members for everything, including: access to physical rooms (or virtual spaces) to host meetings; knowledge and insights to support discussions; drafts of potential guidance text; review and proofreading of guidance documents; and funding hardcopy distribution. It is a great credit to the SASWG Chair, the volunteers' employers and, most especially, the volunteers themselves that the SASWG has functioned so well and achieved so much.

The Challenge

Many aspects of systems engineering are equally applicable to autonomous systems and traditional (or "non-autonomous") systems. The same holds for, on the one hand, conventional software engineering and, on the other, AI and ML. Despite these similarities, autonomous systems, and the associated enabling technologies, present unique safety assurance challenges.

The requirements placed on autonomous systems can be difficult to verify. Top-level requirements can be stated, for example, "Do not run over pedestrians". However, since autonomous systems are used in open, dynamic environments, these requirements cannot be decomposed to a simple set of behaviours in a small number of situations. If that were possible then there would be no need to invoke the complexity associated with autonomous systems.



A similar discussion applies at the computation level. One of the key principles of software safety assurance assumes a traceable, hierarchical decomposition of requirements that encapsulate the safety concerns [3].

Typically, ML approaches do not include this type of decomposition. All the requirements are captured abstractly, for example, via the training data, model (e.g. neural network) structure and training approach (e.g. loss function).

Another challenge in providing safety assurance for autonomous systems is the pace of technological development. Whilst this challenge is not unique to autonomous systems, the significant amount of investment in this area, combined with the vast quantity of academic research that is being published (especially in relation to ML), makes it difficult to identify Recognised Good Practice (RGP). Another consequence of this rapid development is that the best performing technologies are somewhat immature, especially from a safety perspective. This means that rare, but important, failure cases may not have been identified. This also applies to security vulnerabilities, which could have significant safety implications.

The Scope

SCSC-153A aims "*to provide clear, practical, pan-domain guidance on the safety assurance of autonomous systems*". This is clarified in statements describing the document's scope. For example:

- There is a deliberate focus on aspects directly related to autonomy, and enabling technologies such as AI and ML, rather than more general safety engineering or system engineering, where it is assumed that relevant general standards, guidelines and best practice will be applied. The intent is to avoid duplicating existing guidance relating to these general topics.
- There is a deliberate focus on AS [Autonomous Systems] that use AI developed using ML. Although it is possible to envisage AS that do not use these technologies, AI and ML are considered to represent the greatest assurance challenges; they are also expected to be widely used.

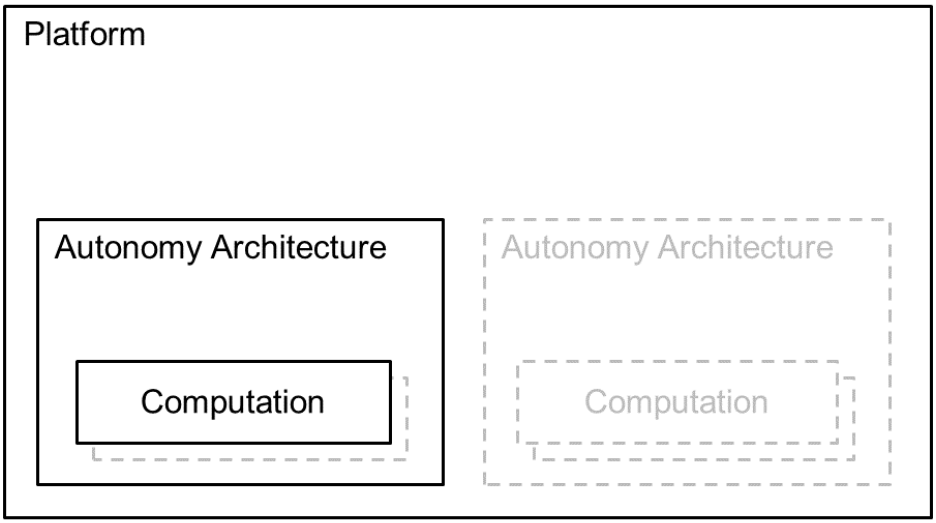
These reflect the SASWG's intent to focus on things specifically related to autonomy. This has allowed helpful guidance to be produced in an efficient manner. It does mean that SCSC-153A should be used alongside other guidance and standards.

The Structure

When SCSC-153A was being developed, there was no accepted way of structuring considerations (safety or otherwise) related to autonomous systems. Consequently, the group developed its own approach. This is based on three "frameworks". As SCSC-153A states:

- The computation-level framework addresses implementation at the software and computational hardware levels. It focuses on mapping an input to an output. Activities associated with this level typically relate to fault prevention. This is the lowest conceptual level considered.
- The autonomy architecture-level framework addresses how computations can be integrated into a system, or platform. Activities at this level typically relate to fault tolerance.
- The platform-level framework addresses what the final autonomous entity should do and what effects it should have on its environment. In essence, the focus is on requirements. This is the highest conceptual level considered.

As illustrated below, a platform may contain multiple "autonomy architectures". Likewise, an autonomy architecture may contain multiple computations.



To illustrate these concepts, SCSC-153A includes two examples. For ease of reference, these are reproduced below.

	Illustration One	Illustration Two
Example Platform	Self-Driving Car	Medical Diagnosis Application
Example Autonomy Architecture Components	Sensor Health Checks, Sanity Checks on Generated Route	Integrity Checks on Supplied Image, Using Multiple Classifiers
Example Computation	Route Planning	Image Classification (Benign / Malignant)

Each framework has an associated set of projections. These were used to help identify specific objectives. Whilst they proved useful, they also caused some confusion. In essence, "*projections provide different perspectives; they provide different ways of viewing each framework*". Hence, they share some characteristics with top-down and side-on projections of the same physical object. Like those projections, the projections in SCSC-153A "*are not intended to be strictly independent, distinct or non-overlapping*". For completeness, the projections are outlined in the following table.

Framework	Projection	Outline
Computation	Experience	Focused on the data that is available to train (or develop) the algorithm
	Task	Focused on the performance of the implemented algorithm; emphasizes requirements
	Algorithm	Focused on the type of algorithm that is used; emphasizes implementation
	Software	Focused on the software used to develop the algorithm and, separately, support its operational use
	Hardware	Focused on the computational hardware that is used, both for development and for operational use
Autonomy Architecture	Tolerance	Focused on tolerating faults and failures in computation-related items
	Information Provision	Focused on recording and maintaining information for subsequent use
	Adaptation	Focused on how updates to the algorithm (and associated software and hardware) are managed
Platform	Behavioural Specification	Focused on platform-level specification; includes defining the scope of autonomous aspects
	Interacting Items	Focused on things intended or required to interact with the platform, not directly owned by the platform developer or operator
	People	Focused on how the platform interacts with people; adopts a whole lifecycle perspective
	Environment	Focused on things in the operational environment that are outside the control of the developer or operator

The Objectives

There are three parts to each objective in SCSC-153A:

- The objective's text;
- A discussion illustrating how the objective contributes to autonomous system safety;
- Examples of approaches that could be taken to satisfy, or partially satisfy, the objective.

The third of these is, perhaps, the most interesting. In developing SCSC-153A, the SASWG was clear that the document should not contain objectives that were impossible to achieve. There was also a desire to avoid being overly prescriptive and an acknowledgement that the "state of the art" was subject to rapid change. The result is a collection of examples that draw heavily on academic papers: SCSC-153A includes 88 references. Moreover, SCSC-153A clearly states: "*These examples are not intended to be prescriptive; there may be other ways of satisfying an objective. Likewise, the examples do not necessarily represent a preferred way of satisfying an objective.*"

It is hoped that the separation of discussion (why the objective is important) from examples (how the objective might be satisfied) will help address the twin difficulties of, firstly, guidance material lagging excessively behind technological developments and, secondly, guidance material changing regularly [4]. In particular, an ambition is for the majority of future changes to SCSC-153A to be located in the examples. In total, SCSC-153A includes 45 objectives. The allocation of these to different framework levels is shown below.

Framework Level	Number of Objectives
Computation	19
Autonomy Architecture	11
Platform	15
Total	**45**

The Justifications

The SASWG devoted considerable effort to justifying the contents of SCSC-153A. Detail is provided in a series of appendixes. More specifically:

- The projections used in the computation framework were compared with a number of broadly similar items, including ones produced by the safety community [5] and ones developed from an ML perspective [6]. These projections were also compared against typical activities associated with, firstly, software development and, secondly, ML.
- The objectives listed in the computation framework were compared with a historical paper on neural network safety [7]: this provides an enduring perspective, rather than one dominated by recent technology trends. They were also compared against an analysis of gaps identified by viewing an automotive standard from an ML perspective [8].
- The projections used in the platform framework were compared with other similar categorisations, including ones developed by US authorities [9], academia [10] and industry [11].
- The complete set of SCSC-153A objectives was compared with the structure of the Assuring Autonomy International Programme (AAIP) Body Of Knowledge (BOK) [12]. The objectives were also compared with the sections in "*The Standard for Safety for the Evaluation of Autonomous Products*", UL4600 [13].

This collection of activities provides confidence that the SCSC-153A objectives are necessary and sufficient. It is, however, important to remember that autonomous systems, AI and ML are areas of rapid technological development. That, inevitably, constrains the amount of confidence that can be placed in any guidance document relating to this area.

The Relations

As noted above, since the SASWG began its work, two related items have been developed, specifically the AAIP BOK and UL4600.

The AAIP is funded by Lloyd's Register Foundation and the University of York. It is addressing global, pan-domain challenges associated with the assurance and regulation of Robotics and Autonomous Systems (RAS). The AAIP BOK is intended to become a key reference source for information related to these topics. It consists of a structured set of assurance objectives, each of which could be required for an assurance case.

UL4600 has been developed by Underwriters Laboratories (UL) and Edge Case Research (ECR). Despite the general terms in its title, UL4600 is targeted at the automotive domain. It concentrates on ensuring that a valid safety case is created. This would be expected to comprise goals, argumentation, and evidence.

The following table summarises the main properties of the AAIP BOK, UL4600 and SCSC-153A.

	AAIP BOK	UL4600	SCSC-153A
Domain	Pan-domain	Automotive focus	Pan-domain
Scope	System-level, based on a Sense-Understand-Decide-Act-like architecture	Whole system view, including systems engineering	Tight focus on autonomy-enabling technologies
Structure	Decomposition-based	Standard, with explicit clauses to produce arguments	Frameworks, projections and objectives
Approx. Size	67 assurance objectives	~190 clauses	45 objectives

All three items are concerned with autonomous system safety. However, they differ in their scope and utility. For example, the AAIP BOK's use of information from practical demonstrators and its non-linear nature may mean it offers most value by providing specific guidance for particular challenges. Conversely, the detailed nature of UL4600 may mean it is well suited to delivery of a mature product, intended for general use, most notably in the automotive domain. Alternatively, the smaller, more accessible nature of SCSC-153A may mean it is a valuable resource for early-stage or specialist products. In addition, the "autonomy architecture" framework provides a neat separation between platform and autonomy-enabled sub-systems, allowing SCSC-153A to be used at the sub-system level.

The smaller, more accessible nature of SCSC-153A may mean it is a valuable resource for early-stage or specialist products.

Given this discussion, it is suggested that the AAIP BOK, UL4600 and SCSC-153A be viewed as complementary (rather than as in competition).

The Future

After two years of rapid change and growth, there are no plans for a new edition of "*Safety Assurance Objectives for Autonomous Systems*" to coincide with SSS'21. Instead, SCSC-153A is being applied to a small number of specific cases. Experience gained from these applications will be used to refine the document, as necessary. Likewise, in due course, the document's content will be revised to account for developments in technology and safety standards.

SCSC-153A is a useful document and the SASWG has already accomplished a lot. However, much work remains to be done. Support in this endeavour would be greatly appreciated, whether that comes in the form of joining the SASWG, providing comments on SCSC-153A, or in any other way.

References

[1] SASWG, Safety Assurance Objectives for Autonomous Systems, v1.0, Feb 2019, scsc.uk/scsc-153

[2] SASWG, Safety Assurance Objectives for Autonomous Systems, v2.0, Feb 2020, scsc.uk/scsc-153A

[3] R. Hawkins, I. Habli, and T. Kelly, The principles of software safety assurance, 31st International System Safety Conference, Boston, Massachusetts USA, (2013).

[4] C. Johnson, Role of regulators in safeguarding the interface between autonomous systems and the general public, (2016).

[5] M. Douthwaite and T. Kelly, Safety-critical software and safety-critical artificial intelligence: Integrating new practices and new safety concerns for AI systems, in Evolution of System Safety, Proceedings of the Twenty-sixth Safety-Critical Systems Symposium, Safety-Critical Systems Club, 2018. ISBN 978-1979733618.

[6] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, What's your ML test score? A rubric for ML production systems, in NIPS Workshop on Reliable Machine Learning in the Wild, 2016.

[7] D. Bedford, G. Morgan, and J. Austin, Requirements for a standard certifying the use of artificial neural networks in safety critical applications, in Proceedings of the international conference on artificial neural networks, 1996.

[8] R. Salay and K. Czarnecki, Using machine learning safely in automotive software: An assessment and adaption of software process requirements in iso 26262, arXiv, 1808.01614 (2018).

[9] Automated Driving Systems 2.0: A Vision for Safety, published by the US National Highway Traffic Safety Administration (NHTSA).

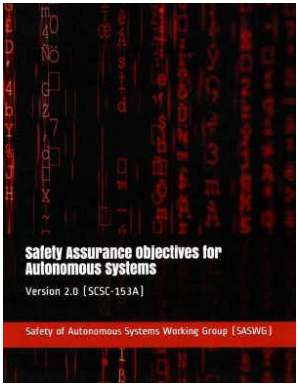
[10] AI Safety Landscape: https://www.ai-safety.org/landscape

[11] Uber Advanced Technologies Group, Safety Case Framework: https://www.uber.com/us/en/atg/safety/safety-case-framework

[12] AAIP Body Of Knowledge: https://www.york.ac.uk/assuring-autonomy/body-of-knowledge

[13] Underwriters Laboratories. UL 4600: Standard for Safety for the Evaluation of Autonomous Products.

Image Attributions
intro image: 173237459 © Andrei Dzemidzenka | Dreamstime.com
2nd image: 132994796 © Akarat Phasura | Dreamstime.com



SCSC Safety of Autonomous Systems Working Group (SASWG)

The Safety of Autonomous Systems Working Group (SASWG) aims to produce clear guidance on how autonomous systems and autonomy technologies should be managed in a safety related context, in a way that reflects emerging best practice.

Comments or suggestions for the Safety of Autonomous Systems Working Group (SASWG) and the Guidance Document are welcome at:

saswg-comments@scsc.uk