STRUCTURING SAFETY CASES FOR AUTONOMOUS SYSTEMS

R.D. Alexander*, N.J. Herbert†, T.P. Kelly*

*University of York, UK, robert.alexander@cs.york.ac.uk, tim.kelly@cs.york.ac.uk
†BAE Systems Military Air Solutions, UK, Nicola.Herbert@baesystems.com

Keywords: autonomous, unmanned, certification, UAVs.

Abstract

Def Stan 00-56 requires a safety case to be built before an autonomous system can be certified, but there is no current guidance on how such a case should be structured. The authors have reviewed several plausible approaches to structuring a safety case, including arguing human equivalence, deriving necessary capabilities from a Level of Autonomy scheme, and by deriving an explicit rationale for the Unmanned Systems Safety Guide recently published by the US Department of Defense. From this, we have produced an initial recommended approach. The process of deriving it has revealed that much of the published advice on autonomous system safety is either of very low value or potentially dangerous.

1 Introduction

There are genuine safety fears about the safety of Autonomous Systems (AS). There have already been accidents in which unmanned aircraft have caused property damage and come close to causing deaths – one example is reported by Johnson and Shea in [1].

Fortunately, for the UK military case, there is already a clear requirement to argue the safety of new equipment item – the authors noted in [2] that Def Stan 00-56 [3] requires a safety case to be built before an autonomous system (AS) can be certified. There is, however, no current guidance on how such a case should be structured. There are several promising approaches, and this paper explores them with a view to proposing a single combined method.

2 Review of Possible Approaches

2.1 Human Equivalence

The basis of this argument is that an AS is acceptably safe for use in a safety-critical environment because the AS is at least as safe as an equivalent system operated by a human.

The requirements currently placed on humans in a given role are the most fruitful source of safety requirements for an AS that will perform that same role. There are a variety of sources that can provide such information (such as rules and laws, training manuals, operator guidelines and standard emergency procedures).

There is pressure for arguing human equivalence as the primary means of achieving AS safety in some domains. The Civil Aviation Authority states in [4] that an Unmanned Air Vehicle (UAV) should be indistinguishable from a manned aircraft, to the extent of using voice communications with Air Traffic Control (ATC). This introduces new safety-critical functionality, most likely a voice communications link from operator to UAV (and hence to ATC). If this extra requirement cannot be met reliably, then the UAV will need to use speech interpretation technology, which is wholly unacceptable from a risk perspective.

According to Godwin [5], a UAV, like a manned aircraft, must behave appropriately when flying in the vicinity of an aerodrome. An air vehicle should be able to recognise that they are in the vicinity of an aerodrome, recognise the pattern of air traffic around the aerodrome and conform to it so as not to cause a hazard to the other air traffic. This includes the requirement to turn towards the left *unless ground signals indicate otherwise*. This introduces the need for more new safety-critical functionality, the ability of the UAV to visually recognise other aircraft and signals on the ground.

In order to claim human equivalence, a UAV must therefore provide a wide range of capabilities, and must be able to perform those capabilities to the same level of integrity as an equivalent human-operated system. The requirement for these kinds of capability is not specific to UAV platforms and equally applies to AS in other domains

It has been suggested that existing Aircrew Standard Operating Procedures (SOPS) could be used to derive instructions for AS to carry out. SOPS encapsulate the expert knowledge of pilots in a set of explicit operating rules and procedures. However, Wright presents an example in [6], of a situation where an airline was operating with SOPS that the aircrew and training staff knew to be inaccurate.

Wright found that pilots annotated their own copy of the pilot operating procedures in their ‘quick reference handbook’, and circulated an unofficial but more comprehensive version of the handbook based on their own experience and what they had learned from simulator training. A UAV relying on SOPS would not have access to this kind of undocumented

knowledge. This demonstrates, therefore, that official rules and procedures are never entirely complete sources of knowledge for deriving safety requirements. Other sources need to be explored if the actual criteria for human equivalence are to be derived.

A pilot's knowledge and skill is often exhibited when they are faced with novel, unfamiliar or emergency situations, for which no written procedures are available to help them resolve a situation safely. There have been a number of examples where pilots have demonstrated great skill and inventiveness in averting or mitigating an accident, such as the Sioux City Accident in 1989 [7] and the Gimli Glider Accident [8].

However, despite these skills, human operators are also very susceptible to making errors in monitoring and cross checking activities, especially at times when there is a high operator workload. It is also well known that humans are vulnerable to stress, fatigue, boredom and absent-mindedness [9].

It is often assumed that manned aircraft are inherently safer than AS. Unexpected pilot skill or knowledge can give a safer outcome from an incident, but there are also that some aspects of human behaviour that can contribute to accidents. It follows then that AS provide opportunities to *increase* safety, provided they are not unduly constrained by the need for arbitrary human equivalence.

2.2 OODA Decomposition

There are already safety case patterns for breaking down a case by an architectural model (such as the Control System Architecture Breakdown argument presented by Kelly in [10] and the software/hardware contribution argument presented by Weaver in [11]). The Observe-Orient-Decide-Act model (see Boyd in [12]) can be used to represent AS behaviour, and therefore provide a basis for structuring and AS safety case.

This provides a solid basic structure, as an alternative to a more conventional hazard-based breakdown, and allows us to build in explicit knowledge of the AS architecture we are using. In itself, however, it doesn't provide much of a guide for deriving the specific safety goals that will make up the lower levels of the safety case.

2.3 Manual Supervision

Manual supervision of AS is often proposed as a panacea for safety issues. The core of such an argument is that an AS can be monitored remotely by a human supervisor. The supervisor is able to intervene in the event of a hazardous situation and take appropriate action to ensure that no harm occurs.

There is already some evidence of human factors playing a role in a UAV accident [1]. In April 2006, a Predator UAV crashed in Arizona. This accident was caused by a UAV operator switching control of the UAV from one operator position to another position when he discovered that the first position had 'locked up'. When this change was made, the operator forgot to alter the controls on the second console to match the controls on the failed console. This error resulted in the UAV's engine fuel shut-off valve being commanded closed, which starved the engine of fuel.

When making a manual supervision argument, reliance is placed on the human supervisor to detect the existence of a hazardous situation. In order to detect a hazardous situation and make an adequate assessment of a situation, an operator must have sufficient, timely and correct information about the AS and its surrounding environment. The correct perception of the environment is known as *situational awareness*, and it is noted in [13] as being a key safety concern when making this kind of argument. The quality, quantity and source of the information that is presented to the supervisor affects the controller's perception of events, and the controller's ability to accurately diagnose a problem and decide on the correct action to be taken.

The supervisor's ability to accurately detect a problem is critical to making the supervision argument because if a hazardous situation occurs and the supervisor does not become aware, then the operator cannot prevent an accident. This situation is effectively the same thing as having no supervisor at all. A developer must be able to argue that the supervisor can be relied upon to mitigate against all hazardous situations that might arise.

Because the supervisor will largely rely on the information displayed to them by the AS control system, it is also important that the user interface used by the supervisor provides the correct information, and enough of it, in a form that the supervisor can usefully interpret under a realistic workload. In many cases, this information will come from the AS via a communications link. This link must, therefore, be reliable and available enough that the supervisor is continuously updated with the current situation.

In addition, the supervisor must have sufficient skills and knowledge about the information and the interface to be able to understand what the information presented is telling them about the AS and its environment.

Having the ability to make this argument is an important tool, but its effectiveness must be assessed on a case-by-case basis. Given the issues involved, especially human factors issues with AS that have high autonomy and short response times, this argument is not an automatic solution.

2.4 According to Level of Autonomy

A variety of Level of Autonomy (LoA) schemes have been proposed (the most well-known is probably Clough in [14]). It has been suggested that the risk posed by an AS will be determined by its LoA. If that is true, then one could potentially use the SIL (Safety Integrity Level) scheme used by Def Stan 00-56 Issue 2 [15] and IEC 61508[16]. Such a

scheme relates the safety-criticality of a system or component to the level rigour required in its development process.

There are two problems with such an approach. First, it is true that each increase in LoA introduces new capabilities that will have implications for safety, but such capabilities may also prevent certain hazards occurring. Increasing LoA may increase or reduce risk; most likely it will do both in different respects.

Some examples of this first problem can be drawn from the Clough LoA scheme. At Level 1, the 'Observe' category has *"Preloaded mission data"*. This may *increase* safety because the AS can continue its mission if it loses communication with its operator, but there is the risk that the operator could load data that is mismatched to the situation. For example, the operator might load a map that is out of date, or that is expressed in mis-calibrated coordinate axes. At Level 2, the same category has *"Health/status sensors"*. It is obvious that this can increase safety through early warning of a hazardous degradation, but it also presents the risk of an incorrect *positive* health report, which might cause the operator to ignore other evidence of problems (e.g. observations of erratic movement).

The second problem is that SIL-based safety arguments rely on claims about development processes, and in 00-56 Issue 4 process evidence is relegated to low-criticality systems or claims. Part 1 of the standard states *"Within the Safety Case, the Contractor shall provide compelling evidence that safety requirements have been met. Where possible, objective, analytical evidence shall be provided"*, and Part 2 expands this by noting that *"In general, arguments based on explicit, objective evidence are more compelling than those that appeal to judgement or custom and practice."*

LoA schemes, and degree of human supervision, are not really separate concepts. Rather, any LoA defines a degree of supervision. As noted in Section 2.3, a claim that supervision leads to safety isn't automatically acceptable; rather, each such claim must be supported by argument and evidence.

2.5 According to DoD UMS Precepts

The US Department of Defence has clearly expended great effort on producing their Unmanned Systems Safety Guide for DoD Acquisition [13], but the resulting document has very little new content. Much of it is generic, rather than Unmanned Systems (UMS) specific, definitions of key terms are unclear and the "precepts" that form the core of it are vague.

The guide is structured around a set of 'Top-Level Mishaps" (TLMs), which identify events and situations that are to be avoided, along with a set of precepts that should be followed in UMS project management, design and operation. The guide's implicit claim is that if the precepts are diligently followed then the TLMs will not occur.

The guide describes a TLM as *"a mishap outcome that can be caused by one or more hazards"*. Nine TLMs are listed in the guide. They are a mixture of accident descriptions (e.g. TLM-4 – *"Self-damage of own system from weapon fire/release"*), consequences of accidents (e.g. TLM 5 – *"Personnel injury"*) and hazards that could give rise to accidents (e.g. TLM 1 – *"Unintended/Abnormal system mobility operation"*). Combining qualitatively distinct entities at the top level makes it hard to provide a systematic way of resolving these entities, or even to assess whether we've achieved such a systematic handling.

Viewed in the context of a UMS-specific safety guide, some of the individual precepts are extremely poor. Probably the worst example is DSP-11 *"The UMS shall be designed to minimize the use of hazardous and toxic materials"*. As a general precept for equipment safety, this is sound and necessary. In a safety standard ostensibly dealing with the safety of a novel, high-risk class of systems, it has no place at all. It actually serves as a distraction from the UMS-specific safety guidance. The entire precept is completely generic; there is nothing in its body text that is specific to UMS. Many of the precepts are weak in this way.

Those precepts that *are* highly relevant are underdeveloped. For example, OSP-1 states that *"The controlling entity(ies) of the UMS should have adequate mission information to support safe operations"*. The body of the precept gives examples of information that the operator may need (e.g. mission objectives, CONOPS, weather conditions). Here, however, it is the detail that matters – the definition of 'adequate' will depend heavily on the system and the situation. Stating that 'adequate' information is needed is of little value – instead, developers need hard criteria for adequacy and a way to derive the information requirements in a given instance.

The guide lacks a well-developed argument for the completeness of the TLMs or the precepts. All that it offers in this regard is a brief overview of the process used in their development (this is repeated in [17]). For the precepts to be directly useful in 00-56 certification such an argument would be to be created. It may be possible to derive an explicit rationale for the precepts (by studying the explanatory text and other publications related to the guide), and thereby build a safety case, but it is likely that such a case would have many gaps.

3 Our Method

Reviewing the approaches to safety case structure discussed in the previous section, we suggest that an effective approach to AS safety analysis will involve:

- A rich definition of the operational scenarios and context
- A set of capabilities derived from the scenarios
- A capability-based hazard analysis

The expected operational scenarios that the AS will be involved in must be documented. This defines the context (in terms of mission, terrain, peer ASs, and interaction with humans) that the AS will operate in, and leads directly to the hazards that will be present and therefore to the final safety requirements. It is critical that discussion of scenario and context takes account of *interactions* between the AS and its peers, and between the various roles that the AS has.

The definition of operating scenarios is particularly important because of the set of AS capabilities that they imply. In many cases, the creation of functional and performance requirements for an AS will be already be driven by expected scenarios. It is important that developers also identify safety-specific capabilities; capabilities that are not required for the AS to complete its mission but that are required to avoid an unacceptable risk of harm.

Typical methods of safety case construction require that the set of hazards presented by the system be identified and acceptably managed. For AS, this analysis can proceed using the identified critical capabilities as its central organising principle. Each capability will present a variety of possible hazards, stemming from a failure to provide the capability, an incorrect implementation of the capability, or from unexpected side-effects of employing the capability.

There are a range of sources that can be used to discover the capabilities needed in a particular context. Section 2.1 gives a discussion of human equivalence as a source (using human equivalence for this purpose does *not* require that the overall safety argument be cast in human-equivalence terms). Simulation and prototyping may also be valuable sources in the later stages of AS development. Similarly, LoA schemes may reveal needed capabilities.

To show that an identified hazard has been prevented or mitigated, an argument can be made based on the inherent safety of the ASs algorithms and behaviour, based on human supervision, or based on additional automated safety functions. Figure 1 shows a fragment of a safety argument, where the hazard of a UAV coming too close to another aircraft is argued to pose only an acceptable risk. In the fragment, the probability of an unsafe course being plotted in the first place is claimed to be no worse 1×10^{-3}. In the event that such a course is plotted, it is argued that there is only a 1×10^{-2} chance of the human supervisor failing to spot and correct this error. Should the human fail, it is argued that the UAV's automated response to the TCAS collision avoidance system will fail to prevent a collision only 1 time in 10. Assuming only a single cause for this hazard, this gives a risk of no worse than 1×10^{-6}, which meets the definition of "tolerable" given in the context node.

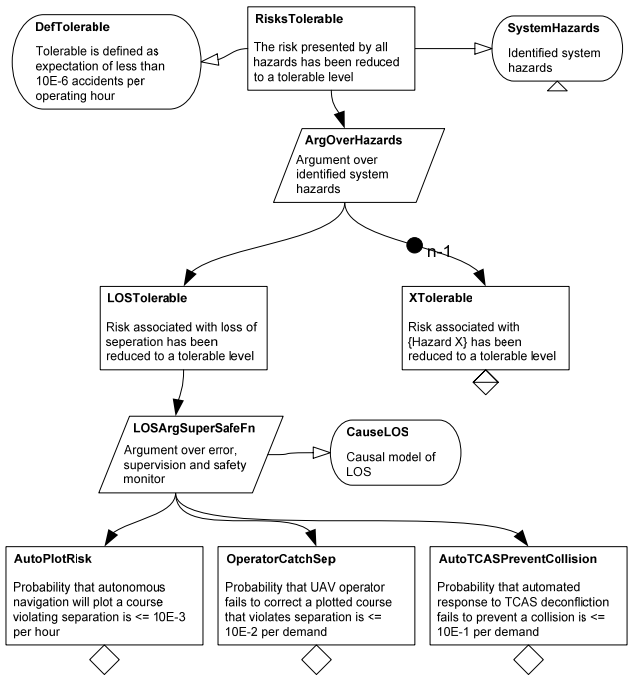

Figure 1 - Safety Argument Fragment

We do not, yet, have a specific method for performing capability-based hazard analysis, or for turning those hazards into specific safety requirements. Producing this will be the focus of our immediate future efforts.

4 Cross-cutting Concerns

There are a number of challenges for AS certification that are largely independent of the safety objectives adopted.

Def Stan 00-56 is subject to the Health and Safety at Work Act, and therefore all safety cases must make the (possibly implicit) claim that they have reduced risk As Low as Reasonably Practicable (ALARP). An ALARP argument is one that claims (a) that the level of risk posed by a system is basically tolerable and (b) that further risk reduction would disproportionate costly. This is a problem because some approaches to arguing safety (such as the CAA's insistence on voice communication – see Section 2.1) may not achieve an ALARP claim.

Regarding the evidence used in a safety case, 00-56 states that "*The quantity and quality of the evidence shall be commensurate with the potential risk posed by the system, the complexity of the system and the unfamiliarity of the circumstances involved.*" The quality of evidence is often referred to as its *trustworthiness*, and this leads to a level of *assurance* of the overall safety of the system. The high risk posed by many AS, along with their inherent complexity and obvious novelty, means that we will need to make high assurance safety claims. This will lead to a need for highly trustworthy evidence, or to arguments that combine multiple sources of independent evidence so as to provide the necessary assurance.

The core of safety analysis is prediction of what hazardous actions a system may perform during its lifetime. Predictability is inherently problematic for AS in three ways: at design time, as part of certification activity, and for peers of the AS during operation:

- At design time – AS developers need to know, ahead of time, what the AS should do in certain dangerous situations. This is not easy – for example, Wright's work on aircraft emergency procedures (discussed in Section 2.1) shows that aircraft manufacturers don't always know what actions should be taken in (expected) emergencies.
- When certifying – Often, a safety analyst will ask questions such as *"What if we lose contact with the AS after authorising weapons use but before receiving confirmation of receipt?"*. Many of the novel control technologies proposed for AS (such as learning or planning algorithms) make it hard to answer these kinds of questions.
- During operation – Peers include other AS, the operators of manned vehicles, and humans who are not part of any vehicle. In military situations, they include both explicit allies and neutral third parties. All of these peers need to predict what the AS will do in order to make their own behaviour decisions and maintain safety.

The combination of a need for high assurance of safety with the above difficulties in prediction make AS inherently difficult to certify. AS are typically proposed for use when we *can't* predict what situations will be encountered, and hence when we *can't* know ahead of time exactly what behaviour will be needed. In operating AS at all, therefore, we sacrifice some predictability. If advanced forms of AS are to be operated at all, we will have to find ways to bound this unpredictability while retaining their valuable flexibility.

5 Conclusions

Based on a review of commonly proposed approaches, we have established a way forward for certifying AS. If a safety engineer adopts the safety objectives given in this paper, and conducts a thorough hazard analysis, then the inherent safety challenges of any individual AS concept will become apparent. These individual challenges can then be dealt with. Failure to do this may lead to operators taking on unacceptable risks.

Specific avenues for further work have been identified: principal among these is a general capability-based hazard analysis method, along with a range of safety analysis techniques for specific technologies. Beyond this, there are several cross-cutting challenges related to assurance and predictability that will be need to be overcome if the most advanced AS concepts are to become a reality.

Acknowledgements

The work reported in this paper was funded by the Systems Engineering for Autonomous Systems (SEAS) Defence Technology Centre established by the UK Ministry of Defence. We would like to acknowledge the help and support of Andrew Miller (BAE Systems), Richard Hawkins (University of York) and Martin Hall-May (University of York) in this research.

References

[1] C. W. Johnson, C. Shea, "The Hidden Human Factors in Unmanned Aerial Vehicles," in *Proceedings Of the 26th International Conference on Systems Safety*, Vancouver, Canada, (2008).

[2] R. Alexander, M. Hall-May, T. Kelly, "Certification of Autonomous Systems under UK Military Safety Standards," in *Proceedings Of The 25th International System Safety Conference (ISSC '07)*, (2007).

[3] "MoD Interim Defence Standard 00-56 Issue 4 - Safety Management Requirements for Defence Systems," Ministry of Defence, (2007).

[4] "CAP 722—Unmanned Aerial Vehicle Operations in UK Airspace: Guidance.," UK Civil Aviation Authority, (2004).

[5] P. D. Godwin, *The Air Pilot's Manual Volume 2 - Aviation Law, Flight Rules and Operational Procedures & Meteorology*, 7 ed: Air Pilot Publishing Ltd, (2004).

[6] P. C. Wright, S. Pocock, R. E. Fields, "The prescription and practice of work on the flight deck," in *Proceedings Of The ninth European conference on cognitive ergonomics*, (1998).

[7] "Sioux City NTSB Accident Report," NTSB AAR-90-06, (1990).

[8] M. Williams, "The 156-tonne Gimli Glider," in *Flight Safety Australia*.

[9] D. Beatty, *The Naked Pilot: The Human Factor in Aircraft Accidents*: Airlife Publishing Ltd, (1995).

[10] T. P. Kelly, "Arguing Safety - A Systematic Approach to Managing Safety Cases," PhD Thesis, University of York, (1999).

[11] R. A. Weaver, "The Safety of Software - Constructing and Assuring Arguments," PhD Thesis, University of York, (2003).

[12] J. R. Boyd, "A discourse on winning and losing," Air University Library, Maxwell AFB, Alabama, USA Tech. Rep. MU43947, (1987).

[13] "Unmanned Systems Safety Guide for DoD Acquisition," US Department of Defence, (2007).

[14] B. T. Clough, "Metrics, Schmetrics! How The Heck Do You Determine A UAV's Autonomy Anyway?," in *Proceedings Of The 2002 PerMis Workshop*, NIST, Gaithersburg, MD, (2002).

[15] "Defence Standard 00-56 Issue 2 - Safety Management Requirements for Defence Systems," UK Ministry of Defence, (1996).

[16] "IEC 61508 - Functional safety of electrical/electronic/programmable electronic safety-related systems," International Electrotechnical Commission, Geneva, (2005).

[17] E. W. Kratovil., R. Barnes., C. Ericson., "The Safety of Unmanned Systems: The Process Used to Develop Safety Precepts for Unmanned Systems," in *Proceedings Of Proceedings of the 25th International System Safety Conference (ISSC)*, Baltimore, (2007).