Safety, Reliability and Security Modelling and Methods for Autonomous Systems

Report of the Discussions of Breakout Session

Authors: John Andrews (Session chair), Stein Haugen, Marilia A. Ramos, Christoph A. Thieme

Assessment of Safety, Reliability and Security of autonomous systems

The introduction of autonomy into systems adds new layers of complexity regarding safety, reliability and security (SRS). The main challenges related to methods for SRS assessment concern the following: i) the adequacy of existing methods; ii) the integrated modelling of hardware, human, software and human interaction; iii) self-learning systems; and iv) data requirements.

Risk Assessment

The general goal of risk assessment is to identify hazardous events, prevent their occurrence and mitigate their consequences. A broadly accepted definition of risk is the expected likelihood of a hazardous event combined with the expected consequences. Important questions arise regarding this definition, such as if risk (in the sense of statistically expected loss) is a relevant measure. Further, is risk related to autonomous systems the same as for traditional systems?

The assessment of the likelihood or frequency of events involving autonomous systems is the most challenging part of risk analysis. An additional challenge concerns risk related to software aspects. Methods to investigate hazards resulting from hardware failure and human error are relatively mature; however, the same is not true for software implementation. The assessment must address not only the occurrences of incorrect responses from the software, but also the failure mode that this induces in the system. Yet, most software reliability methodologies focus on the number of bugs remaining in a code, regardless of their effect on the system. In addition, unlike hardware, the historical performance of a software cannot be considered as indicative of future performance. For autonomous systems the problem is compounded by the fact the software can incorporate self-learning and there is no clear rule-based algorithm to examine. This last property may make a systematic evaluation of the potential hazards problematic and hence the risk quantification breaks down.

Regarding consequences, they may be mostly the same as for non-autonomous systems. Exceptions are systems that currently operate manned and that may become unmanned with the introduction of autonomy, e.g., offshore platforms. In this case, the consequence of an accident could potentially be reduced, given that no life of workers would be threatened. Nevertheless, the negative environmental impact would remain the same.

Risk assessment may be adapted for different applications. Traditional hazards, such as fire and collision, are present in existing frameworks and should be included in autonomous systems' risk assessment. However, autonomy includes new hazardous events for which the risk community must investigate the possibility to address and incorporate in the existing frameworks. Threats due to system connectivity, such as cybersecurity, may be challenging to incorporate in traditional risk assessment frameworks. In particular, the frequency of such events may be difficult to define. Hence, a cooperation and exchange of methods and approaches between industries and application areas is highly necessary.

Reliability and availability

For safety critical applications, there is commonly a strategy for component failures to be 'fail safe'. Whilst enhancing safety, this can have a detrimental effect on reliability. Autonomous systems must be reliable over the time of their mission, as there may not be any option for repair during the mission.

Resilience is indicative of how the system can bounce back after a problem has occurred and therefore may also provide a useful measure of system performance. Availability also indicates the ability of a system to return quickly to the working state following a failure. In the maritime sector, it is advantageous to have a high likelihood of completing several missions for the vessel to be returned to a dock with the required maintenance facilities to prepare the vessel for its next sequence of missions. This is a similar concept to the Maintenance Free Operating Periods proposed by the aeronautical industry.

In addition to being resilient, it is beneficial if autonomous systems can include some form of self-repair. The analysis should not solely rely on probabilities or frequencies, since there is a lack of data and a need for make assumptions in all cases.

Security

Security is related to threats from external agents who have the intention to harm the system. Attacks on autonomous systems can exploit some weaknesses that are particular to those systems, and difficult to foreseen. The assessment of resilience strongly correlates with security: can the system operate after a security breach?

In security, one needs to look at the different realms, including human, software, hardware, society. One of the most frequently used attack methods is to "hack" the human since this may be the weakest link. Humans can also be a security barrier, and their effectiveness needs to be assessed.

In addition to using humans as a breach for an attack, a concern regarding security of autonomous systems are cyberattacks. During a cyberattack, hackers first scan the system and find an open "port" or a vulnerability. They attempt to get the credentials to infiltrate the system. A challenge concerning this type of attack is that while the hacker needs only one port in, the defender must defend all ports. It is therefore necessary to identify which ports are insufficiently protected. A probabilistic method may help in this identification. Also, the analysis of security can leverage from other application areas.

One approach to security analysis is to use game theory. Other methods are attack trees, expert judgements and scenario roleplays. Vulnerability analysis should be a part of security analysis. This is an essential part of security risk assessment in several industries.

Adequacy of current modelling approaches

One of the key topics in the discussion of SRS assessment for autonomous systems is the adequacy of the existing modelling approaches. Could existing approaches be applied directly? The difficulty in obtaining frequencies for some of the events would be an issue and so existing modelling frameworks can work for part of the system assessment only. However, since these methods have served well in the past, and are relatively efficient for existing systems, they should not be dismissed for autonomous systems. Rather, they can be enhanced with additional assessments for the command and control structures.

The types of accidents that can occur in autonomous systems may not be different from conventional systems; yet, the causes to the accidents will change. Autonomous ships, for instance, differ from a traditional ship mainly regarding the responsible agent for decision: with autonomy, some or all the decision-making processes are moved from a human to software. Also, system design and maintenance for autonomous systems may not be direct "copies" of their non-autonomous predecessor (e.g., additional redundancy, predictive or preventive maintenance may be needed to ensure adequate mission reliability).

Qualitative assessment methods are believed to be largely applicable, although improved methods to move further "to the left in the bowtie" are required, with a focus on the causal analysis. Security and “software failures”, however, pose significant challenges, as stated in the previous section. The challenges are essentially related to two aspects: (i) failure to identify all the circumstances that the software need to be able to handle, and (ii) failure to understand how the software works in all circumstances. While the first challenge is related to hazard identification methods, the second is closely related to the self-learning aspects of the software, which provide challenges in verification and testing.

The identification of all the hazards, situations and scenarios that the systems need to deal with is critical. In the car industry, it is attempted to define each subpart of the driving process, for example parking. The problem is then limited to only some parts of the operation, for which safety issues are identified.

To identify all the different events that can happen, the analyst must consider an appropriate level of abstraction for the problem, in addition to historical data and experience. The autonomous platform cannot be considered in isolation. The response to an unsafe state is dependent upon the location and environmental conditions. A car, for instance, will operate in different regions, that may have very different traffic patterns (e.g., in Norway / Sweden or in India / Pakistan). As a validation approach, in the car industry, autonomous systems are running in the background while a human driver is controlling the car. This helps to identify new scenarios that the autonomy should react to, but it does not ensure all possible scenarios are covered.

A risk model needs to accommodate the environment, the weather, and the mode of operation. A challenge is the identification of all circumstances that the system may meet. More complex systems may imply that more systematic methods are needed, e.g., to assess the interfaces. Some methods that are currently used and may be applied to autonomous systems include:

- Fault and event tree analysis (FTA & ETA)
- System theoretic process analysis (STPA)
- Functional Resonance Analysis Method (FRAM)
- Simulations

Fault tree and event tree analysis are traditional methods that focus on the graphical representation of the risk analysis. The methods represent events and not every accident is event driven. It may be the circumstances deviating from those expected that lead to an accident.

STPA is a rather new qualitative hazard analysis method. It is a systemic and systematic approach that treats safety as a control problem. It is not limited to component failure, as the more typical risk analysis methods, but it attempts to identify complex interactive scenarios. Software, hardware, humans, organizations, and regulations can be modeled within the same framework. Also, different levels of abstraction may be employed, and it can be used for all system properties (e.g., safety and security). The main disadvantages are the high need for resources, the lack of competency in the industry, lack of prioritization and ranking, and the lack of the right tools for using the process efficiently.

FRAM is designed to qualitatively communicate risk and the complexity of a system. It is less operational, compared to STPA. The disadvantages are the same as for STPA, i.e., resource intensive, lack of competencies, prioritization, and lack of tools for efficient use.

In general, simulation is a very powerful tool, if applied correctly. It may enable analysts to collect a large number of data for different situations and scenarios at a low cost. Simulation may also include the human, operating in the loop, for a holistic assessment approach. Simulation can be an efficient solution to demonstrate efficiency and transparency of the autonomous system capability.

In the future, simulations should be combined with real (on site) testing to prove to society that the system is safe and capable. It is possible to acquire environmental and operational data offline. Moreover, it is also possible to simulate and test the autonomous systems' responses to very rare events.

Modelling of human, hardware, software and interfaces

Existing risk assessment methods tend to focus mainly on the hardware and human elements of the system. Humans will still be an essential part of autonomous systems in the near future. Depending on the Level of Autonomy, humans will need to remotely control the system or supervise it and step in when problems occur. Models are required to evaluate the contributions from the human in failing to achieve a successful recovery from a problem. It is critical to establish what information is required at handover, how the information is provided, and the time frame for handover.

The probability of software failure is required as input to risk quantification. Software failure is different in nature from physical components. Software will fail when circumstances that have not been predicted by the designers occur or when mistakes have been made by the programmer. Occurrences of these failures are not stochastic but deterministic in nature.

An additional vulnerability in software is that resulting from an intended attack, as stated previously. Hackers exploit unknown vulnerabilities in the system and for critical infrastructure such as transport systems, these may be state sponsored. Since the current and future frequency of these attacks is not related to their historical occurrence, it is not possible to evaluate this requirement for a risk study. Decisions on the risk posed through such cyber-attacks cannot therefore be evaluated with a risk framework and alternative approaches are needed.

Assessment of self-learning systems

Self-learning autonomous systems may develop their own "personality". They may learn and adapt to specific people and environmental conditions, events and actions. If the analysis of an autonomous system SRS (particularly for the software elements) will rely in testing, updates which change capabilities will need to be formally assessed.

An example is autonomous car driving systems, which exhibit very high complexity. They need to account for all road junction types, regional driving cultures and individual driver characteristics. To physically test a vehicle for all potential options encountered for global operation is not viable. In these circumstances, testing and validation are only possible using simulators which can replicate the full range of options encountered (including those rarely encountered). Simulators can conduct the testing considerably faster than rear time road testing.

Immaturity of risk assessment and validation methodologies in this area pose a potential safety risk. New methods and tools need to be developed. The implementation of autonomous systems should only proceed at the pace of the assessment methods.

Resilience

Resilience engineering approaches were considered to offer an alternative philosophy to risk assessment by which autonomous systems can be assessed and worthy of more detailed consideration. A resilient system is one which can anticipate, absorb, adapt to or rapidly recovered from a disruptive event. This focus on the ability of the system to recover from an unwanted event gives a means by which software malfunction may be evaluated without the need to predict the occurrence frequency.

When a system fault is observed, a response needs to be fast and the initial incident management may have to be performed without knowledge of its cause. Once the cause is determined a transition from incident management to full system rectification can be implemented. Such an approach enables the system to be safely operated in all circumstances, not only those with a low risk prediction.

The potential benefits of such a resilience approach needs further investigation. Measures of system performance (MoP), which should be predicted throughout any incident, would need to be established. It is expected that these will vary depending on the occurrence of a safety problem or reliability problem. Methodologies to predict how this MoP varied through the phases of threat occurrence, system performance degradation, incident management and full system recovery would be needed and the exact definition of resilience which was predicted from these factors established. Different MoPs would be required for different autonomous system applications.

Real time operational decision support

Conventional risk analysis techniques such as fault tree analysis are usually used off-line in order to certify that a particular system delivers acceptable safety performance. An autonomous system will need to establish when it is no longer operating safely and requires a mission abort strategy to be activated. Determining unacceptable performance can be rule based or it can exploit the system failure analysis approaches in real time to predict when the safety performance is no longer acceptable. Events which represent deteriorated or failed hardware (established through fault diagnostics), changes in environmental or operational conditions can be input as updated event probabilities to the system failure models. The analysis of models formulated as a fault tree can be rapidly performed using Binary Decision Diagrams.

Such approaches have been explored to establish unsafe conditions for pilot-less aircraft, UAVs (Unmanned aerial vehicles), the timeframes in which decisions need to be made would certainly make these approaches applicable in the maritime and marine applications. Since the operating environment of aircraft is less complex than cars, the response time of such predictions may currently limit the potential for automobile application.

Data requirements

Several types of data are needed for the SRS assessment of autonomous systems, including

- Sensor data and understanding of their usage
- Service, repair, warranty and maintenance data
- Experimental data and test data
- Condition data
- Surrogate data gained through simulations
- Data on the frequency and nature of cyber-attacks on the system

These data may be used in real time, in virtual and dynamic models, to manage failures and plan and predict maintenance. However, the data needed is frequently not available and, when available, may be of low. Standards for data collection are needed across companies and sectors.

Some data may be transferred between industries. For example, information related to human factors and human error may be transferred between highly automated systems. It is important that historic data, or data from manned systems, is assessed for their applicability for autonomous systems. Similarly, environmental data needs to be assessed for case relevance.

Data needs to be analyzed together with the associated uncertainty, to ascertain if data is complete or if there gaps in the observations, due to an insufficient monitoring frequency.

Conclusions

From the group discussions the following conclusions were drawn regarding the safety, reliability and security of autonomous systems.

- The software elements of autonomous systems challenge the applicability of current risk assessment approaches. This is due to software malfunction being very different from hardware or human failure and not stochastic in nature. Since their historical occurrence does not indicate future expectations it is not possible to formulate their expected likelihood or frequency. The same problem exists in predicting the frequency of malicious, intentional attacks on the software. Self–learning features of the software also add difficulty in the validation of acceptable performance.
- Many of the currently available methods can still play a part in supporting the safety, reliability and security of autonomous systems.
- New modelling techniques, which holistically capture the strong connectivity and interdependencies between software, hardware and human operators are required.
- Simulations provide a practical approach to assist in the detailed understanding autonomous systems with respect to SRS, to collect data, and to validate systems SRS behavior.
- The concept of resilience engineering is an alternative approach to risk assessment and offers a focus on absorbing and recovering from failure events which can be applied without knowing the frequency of the failure.
- Data requirements are application specific with standards required to ensure quality.

- Quantification methods for SRS can play a bigger part in the future. In addition to the certification process for an autonomous system they can be incorporated for real-time decision support during a mission to identify when the system performance drops below the acceptable threshold.

Group Participants

John Andrews
University of Nottingham, United Kingdom

Jonas Borg
Volvo Penta, Sweden

Kenneth Titlestad
Sopra Steria, Norway

Markus Heimdal
Rolls Royce Marine, Norway

Matthew Minxiang Hu
Haylion Technologies, China

Nikolaos P. Ventikos
National Technical University of Athens, Greece

Odd Ivar Haugen
DNV GL, Norway

Osiris Valdez Banda
Aalto University, Finland

Siv Randi Hjørungnes
Rolls Royce Marine, Norway

Stein Haugen
Department of Marine Technology, NTNU, Norway

Thomas Johansen
Department of Marine Technology, NTNU, Norway

Torgeir Moan
Department of Marine Technology, NTNU, Norway

Yan-Fu Li
Tsinghua University, China