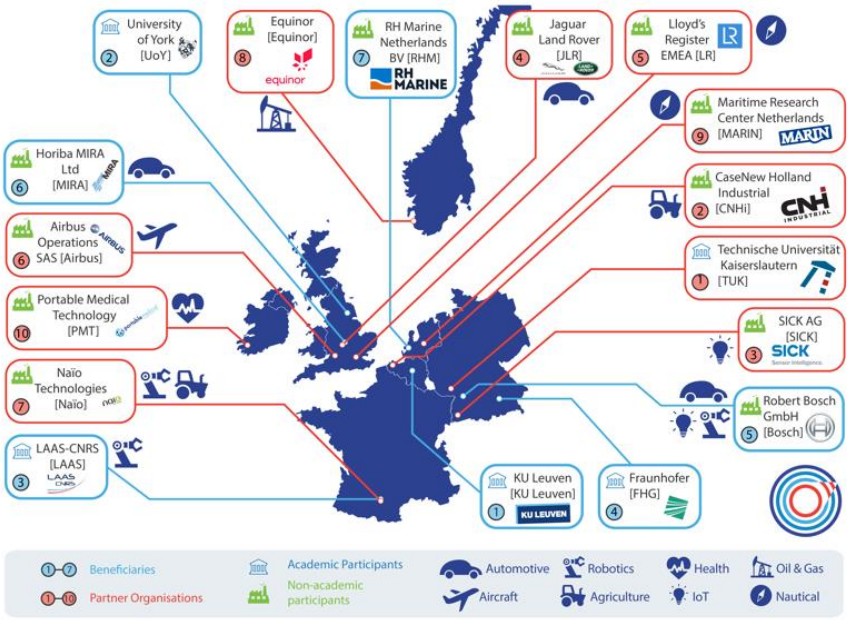
Safer Autonomous Systems Training Network



Few people are really ready to trust autonomous machines and are much less prepared to tolerate a mistake made by them. Davy Pissoort, professor at KU Leuven Bruges Campus in Belgium, discusses the Safer Autonomous Systems Marie-Curie Innovative Training Network – a four-year initiative looking at ways to establish people's trust in autonomous systems by making these systems demonstrably safer.

Introduction

The coming of autonomous systems doesn't just mean self-driving cars. Advances in artificial intelligence will soon mean that we have drones that can deliver medicines, crew-less ships that can navigate safely through busy sea lanes, and all kinds of robots, from warehouse assistants, to search-and-rescue robots, down to machines that can disassemble complex devices like smartphones in order to recycle the critical raw materials they contain.

As long as these autonomous systems stay out of sight, or out of reach, they are readily accepted by people. The rapid and powerful movements of assembly-line robots can be a little ominous, but while these machines are at a distance or inside protective cages we are at ease. However, in the near future we'll be interacting with "cobots" – robots intended to

assist humans in a shared workspace. For this to happen smoothly we need to ensure that the cobots will never accidently harm us.

This question of safety when interacting with humans is paramount. No one worries about a factory full of automated machines that are assembling cars. But if these cars are self-driving, then the question of their safety is raised immediately. People lack trust in autonomous machines and are much less prepared to tolerate a mistake made by one. So even though the widespread introduction of autonomous vehicles would almost eliminate the more-than 20,000 deaths on European roads each year, it will not happen until we can provide the assurance that these systems will be safe and perform as intended. And this is true for just about every autonomous system that brings humans and automated machines into contact.

Autonomous vehicles, indeed all autonomous systems, need to be made safe enough so that people trust them. The destination, therefore, is clear; the route, however, is a difficult one. The Safer Autonomous Systems ITN project is designed to get us to our destination, safely.

Until recently, safety assurance has been integrated into the design processes, based on safety standards and demonstrating compliance during the system's test phases. However, existing standards were developed primarily for human-in-the-loop systems, where a human can step in and take over at any time. They do not extend to autonomous systems, where behaviour is based on pre-defined responses to a particular situation. What's more, current assurance approaches generally assume that once the system is deployed, it will not learn or evolve. On the one hand, advances in machine learning mean that autonomous systems can be given the potential to learn from their mistakes, and the mistakes of all the systems they are connected to, making their abilities to operate safely infinitely better than previous generations. On the other hand, machine learning means more uncertainty about how the system will decide to react to a particular circumstance in the future, making safety assurance a hard task, which can only be accomplished by a highly-skilled, interdisciplinary workforce.

Are you ready yet, to take a seat on an autonomously controlled airplane? If you hesitate to say "yes", then you are tacitly acknowledging the need for a training and research programme such as the *Safer Autonomous Systems (SAS) Innovative Training Network (ITN).*

SAS Objectives

The main objective of the Safer Autonomous Systems (SAS) [1] project – which started on 1st Nov 2018 and runs for four years in total – is to identify ways that we can establish people's trust in autonomous systems by making these systems demonstrably safer. In order to achieve this objective, we identified three challenges to be addressed by the early-stage researchers (ESRs) in their 15 individual research projects. This simply-stated objective, and the interdisciplinary needs required for its realization, is of such complexity that we saw a large training network involving some of Europe's flagship companies – such as Bosch, Airbus and Jaguar Land Rover – together with leading European universities – like KU Leuven and the University of York – as the best way to tackle these challenges, which are briefly described as follows:

- Increased autonomy, by definition, means a significant reduction of the time during which a human is involved in the system's decision making, thereby reducing the residual control afforded to humans. Studies have shown that it may take minutes for a non-actively involved human operator (e.g. a passenger in a self-driving car) to take over control in case of an emergency. Moreover, just putting a self-driving car to a full stop on a busy highway by removing its power (so-called fail-stop behaviour) is definitely not a safe action. In contrast, an autonomous system should be fail-operational (perhaps with reduced functionality) under all circumstances, monitor its own safety and make its own decision about a sensible and safe reaction. The challenge therefore is to design autonomous systems in such a way that they remain safe under all conditions, even in the case of component failures.
- Testing is the most intuitive way to reveal unsafe behaviour. However, autonomous systems must operate in a near infinite range of situations. When we test autonomous systems, we must therefore systematically determine which range and diversity of situations should be simulated and tested. We need to test them on roads, in the rain, and with people in the way. We need to test them when they're in intermittent supervisor contact and when they've got an unbalanced wheel. And we need to test them in all the possible combinations of those cases. Testing autonomous systems in the field is clearly too costly and too time consuming and might even be harmful for the system or its environment. Hence, virtual model-based testing is the only viable option. However, breakthrough solutions are required to guarantee the rigour of our virtual testing and to optimize its overall coverage.
- More autonomy is possible only through new technologies, e.g., machine learning, for which no accepted safety-assurance strategies currently exist. Legacy experience as well as established standards and regulations are lacking. Implicitly or explicitly, current safety-assurance practices and safety standards assume that the behaviour of the system is known at the design stage and can be assessed for its safety prior to system deployment. As autonomous systems might learn and evolve over time, this is no longer possible. This means that meeting the current safety standards for autonomous systems is either impossible to do or completely insufficient to assure safety throughout the life time of the system.

To achieve the main Scientific/Technical (S&T) objective of trust in autonomous systems by overcoming the 3 challenges, we decided on three sub-objectives that are the aims of the project's 3 research Work Packages (WPs):

- **Objective 1:** To integrate guaranteed acceptably safe behaviour directly into the architecture/design of the autonomous system (WP1)
- **Objective 2:** To prove by model-based safety-analysis techniques that the behaviour of an autonomous system remains acceptably safe under all possible conditions (WP2)
- **Objective 3:** To ensure that the safety-assurance strategies that combine the architectural/design measures with the evidence allow us to have trust in the autonomous system, which is very likely to be learning and evolving (WP3).

Overview of the SAS consortium

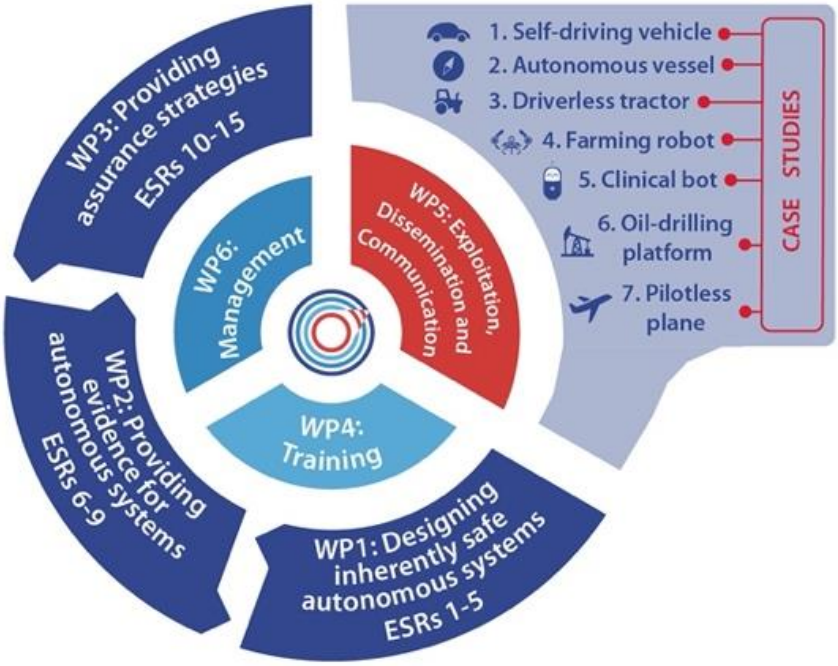
SAS provides specialised training to 15 early-stage researchers (ESRs), for 540 researcher months. The 15 ESR positions are distributed among 7 Beneficiaries in 5 countries. In addition, SAS has 9 non-academic Partner Organisations and one academic Partner Organisation.

The SAS consortium includes the 2 leading members (University of York and Lloyd's Register) of the Assuring Autonomy International Programme (AAIP) – which is the perfect partner for SAS. The AAIP is a prestigious initiative that targets the development of standards, open toolsets and training resources for assurance and the public acceptance of robotics and autonomous systems by collaborating with industry, regulators and research teams around the world.

The already well-established collaborations between the institutions involved will ensure that the network runs smoothly, while strengthening the interactions and the exchange of academic and non-academic resources. SAS aims to actively research the development of safer autonomous systems at multi-nationals like Bosch, medium-sized companies like MIRA and RH Marine, but also to stimulate the development of new safety designs, modelling and assurance techniques by involving the ESRs in SMEs and, potentially, their own start-ups.

SAS Work Programme

The SAS project is based on 6 Work Packages (WPs), three of which are S&T WPs (WP1–3), one for training (WP4), one for Exploitation, Dissemination and Communication (WP5) and one for Management (WP6).



The S&T WPs are organized along 3 research tracks covering the 3 main steps in the safety-assurance process:

- building safety and dynamic risk mitigation into the system by design
- gathering evidence that the behaviour of the system will actually be safe
- combining these into a clear strategy that allows us to put our trust in the system.

WP1: Designing inherently safe autonomous systems

WP1 involves 5 ESRs and tackles the actual safety-aware design, i.e., making safety inherently part of the design process of resilient autonomous systems.

ESR1 and ESR2 take up the challenge of *developing generic frameworks to monitor and handle the safety of autonomous systems during run-time*. The role of such a safety monitoring-and-handling framework is to observe the system and its environment and to trigger interventions that maintain the system's safety. For non-autonomous systems, a human operator takes a significant role in this fault-monitoring and, definitely, in the fault-handling. In contrast, versatile autonomous systems will have to deal with a much richer set of safety rules. Moreover, these safety rules have to take into account the wide application of machine learning in autonomous systems, causing them to evolve over time, and the a-priori largely unknown open-context in which autonomous systems will be applied. ESR1 and ESR2 will be working on related, but complementary projects, with ESR1 having the task to extend current safety-monitoring frameworks such that they cover the whole chain from safety-constraint definition to the actual autonomous reactions to avoid a possible hazard, while ESR2 starts from a MAPE-K cycle (i.e., Monitoring, Analyse, Plan and Execute based on Knowledge) to enable real-time adaptations of functionality, structure, and fault-tolerance mechanisms in order to assure the run-time resilience of autonomous systems.

ESR3 will run in parallel with ESR1 and ESR2 and go one step further and integrate *dynamic safety handling of autonomous systems-of-systems through run-time safety contracts* into the adaptive safety monitoring and handling framework. Driven by trends like ubiquitous computing and cyber-physical systems, new application domains for autonomous systems-of-systems have emerged. In such systems-of-systems, different devices are combined during run-time to fulfil higher-level emergent functionalities in a collaboration that cannot be provided by one of the involved systems on its own. Classic safety assurance relies heavily on a complete understanding of the structure and behaviour, which is not available at design-time for an autonomous system-of-systems. It is therefore more reasonable to use the idea of safety contracts between the different subsystems. Safety contracts are an effective way to conditionally describe the safety guarantees that a component should fulfil in order to make sure that the overall system-of-systems remains safe. Up until now, the use of safety contracts has mainly been limited to static, non-evolving systems. ESR3 will extend this approach to dynamic, modular safety contracts.

ESRs 4 and 5 will work on effective techniques and measures that assure by-design that even under fault conditions the autonomous system remains safe without any human intervention. When autonomy increases, so does the software complexity and thus the likelihood that it contains faults. Therefore, ESR4 focuses on *software design guidelines and testing specifications for non-functional requirements in safety-critical autonomous systems*. Future applications of autonomous systems will rely heavily on different communication technologies to connect and interact with other devices, infrastructure, the "cloud", etc. Although adding connectivity has its benefits, it also adds challenges, among which are most definitely its robustness and resilience. ESR5 focuses on more hardware-oriented design and testing specifications, which make *connectivity work reliably under a diverse range of environments*. This takes into account a combination of stresses, including electromagnetic interference, temperature and vibrations, aging, etc.

WP2: Designing inherently safe autonomous systems

WP2 targets novel methodologies that allow us to evaluate, validate and verify the safety-aware design (WP1), meaning that safety can be guaranteed given the complex environment and extremely varied use-case scenarios that autonomous systems will be subjected to. This challenge cannot be underestimated. As Michael Bolle, President of Bosch, Corporate Research once said in a speech: "*We have looked at what it takes to physically validate autonomous driving, and the time needed was estimated at 100,000 years. We need breakthrough solutions from the research community.*" As physical testing is too costly and too time consuming, we must turn to virtual, i.e., simulation- and model-based, testing.

ESRs 6 and 7 will collaborate to achieve a breakthrough with respect to the overall coverage of the model-based safety analysis. ESR6 will address the issue of the *virtual-worlds generation* and will apply this to autonomous robots. In other words, ESR6 answers the question "which operational situations and environments should be tested in the virtual world?" and starts from a criticality analysis. Once the most critical virtual worlds have been generated, ESR7 will *evaluate and maximize the situation coverage of each of the virtual worlds*. Exploiting techniques like Environmental Survey Hazard Analysis, ESR7 will answer the question about which faults and failure modes should be seeded into the model-based safety analysis to cover the most threatening challenges.

Whereas a classic model-based safety analysis often limits itself to failures of one or multiple components, the open-context nature of autonomous systems forces us to also consider the safety-applications of functional insufficiencies. A typical example being a camera in a self-driving car that should prevent a collision with a human being, but only detects a human correctly in 99.9% of the cases. Therefore, ESR8 is going to look at *model-based system-analysis techniques to determine propagation paths of functional insufficiencies in software-intensive systems* and will use probabilistic ways to model the uncertainties.

Complementary to ESR5 (WP1), ESR9 will also take up the challenge of the strong reliance of autonomous systems on wireless communication and will perform a *model-based system analysis of the robustness of autonomous systems against electromagnetic interference*. Combining efficient statistical electromagnetic modelling with behavioural modelling, the resulting behaviour of an autonomous system upon electromagnetic disturbances will be forecasted and evaluated.

WP3: Providing assurance strategies

The S&T WPs conclude with WP3, which pilots novel safety-assurance strategies, combining the previous 2 research WPs, thereby allowing us to put trust in the safe behaviour of autonomous systems. In total this WP involves 6 ESRs and, besides safety, also covers other design constraints such as security, reliability, availability and liability.

ESRs 10 and 11 both focus on dedicated assurance cases for autonomous systems. Existing standards, processes and practices place a great emphasis on how safety can be certified throughout the design and development stages. However, there is little guidance on how

safety assurance should be maintained throughout the system's operational life. Many assumptions about the environment and the system performance and use, particularly for complex and novel autonomous systems, that are made during the design and development stages might turn out to be incorrect during operation. From a safety point of view, this can threaten the validity of the safety case and weaken confidence in the actual safety of the system. Within SAS, two complementary approaches will be pursued to tackle this:

On the one hand, ESR10 aims at making the transition *from static assurance cases during design-time to executable assurance cases during run-time*. Here, a safety-assurance case is structured argumentation, supported by evidence (WP2), intended to justify that the system is designed (WP1) such that its behaviour is acceptably safe when being put into service. While for non-autonomous systems, the whole safety case is traditionally developed, documented and accepted prior to operation, the safety case for some autonomous systems may instead need to be posed with residual obligations that are only satisfied during run-time. For example, the vast number of possible inter-vehicle and infrastructure configurations that an autonomous vehicle may encounter may require run-time verification of safety properties (such as end-to-end response times, or the integrity of received data) to sustain the safety case. Therefore, ESR10 will establish a new way of working with an executable set of claims that will be sustained and maintained during run-time.

On the other hand, ESR11 will study *assurance-case structures for machine learning in the decision making of highly autonomous systems*. Currently, the use of machine learning or any other artificial intelligence technique, is not recommended for safety-critical tasks. However, many autonomous systems will rely on machine learning and ways to address this are urgently needed.

ESRs 12, 13, and 14 all start from a specific application scenario, i.e., autonomous vessels, clinical conversational bots and autonomous vehicles, respectively. In addition, they always consider other, possibly conflicting, design constraints, as is the case in industrial practice. ESR12 wants *to assure safe autonomous sailing from A to B while minimizing operational costs* by combining a cost-optimization algorithm, a collision-avoidance algorithm and situational awareness. ESR13 looks at the *safety assurance for clinical conversational bots* by combining safety engineering with typical clinical processes. ESR14 will cover the whole *dependability assurance for autonomous vehicles*, covering, besides safety, also reliability, availability and cyber-security.

Last, but certainly not least, ESR15 will take up the emerging challenge of the liability aspects of autonomous systems in safety-critical domains. ESR15 will propose a *liability allocation framework for safe autonomous systems* that explores new avenues for the allocation of liability for autonomous systems in a way that strikes a balance between the commercial interests of operators and manufacturers and the safety and fair compensation of the general public.

Summary

Autonomous systems offer humankind tremendous opportunities, like freeing us from mundane tasks, carrying out risky procedures and generally giving us more time to enjoy the things we like doing. However, we lack trust in many forms of autonomous systems: partly this is human nature, but primarily because these systems, such as self-driving cars, have not demonstrated their safety credentials yet. Only by making these systems safer can we expect their widespread acceptance. The Safer Autonomous Systems (SAS) ETN is about getting people to trust these systems by making the systems safer. In order to achieve this objective and to train a group of highly skilled, responsible, future innovators, we will bring together 15 early-stage researchers (ESRs) to investigate new forms of system-safety engineering, dependability engineering, fault-tolerant and failsafe hardware/software design, model-based safety analysis, safety-assurance case development, cyber-security, as well as legal and ethical aspects. Hence, the most important output of SAS will be 15 well qualified people who have been trained to tackle many of the problems now being faced by European industry.

References

[1] European Training Network for Safer Autonomous Systems, http://etn-sas.eu, accessed 31 Oct 2020

Davy Pissoort, Professor KU Leuven Bruges Campus & SAS Project Coordinator

Davy received his PhD degree in electrical engineering from Ghent University, Belgium in 2005 and worked as an R&D Engineer at Agilent Technologies (now Keysight Technologies) in Ghent. He is currently professor and head of the Mechatronics Group at KU Leuven Bruges Campus.

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No 812.788 (MSCA-ETN SAS).

This article reflects only the authors' view, exempting the European Union from any liability. Project website: http://etn-sas.eu/