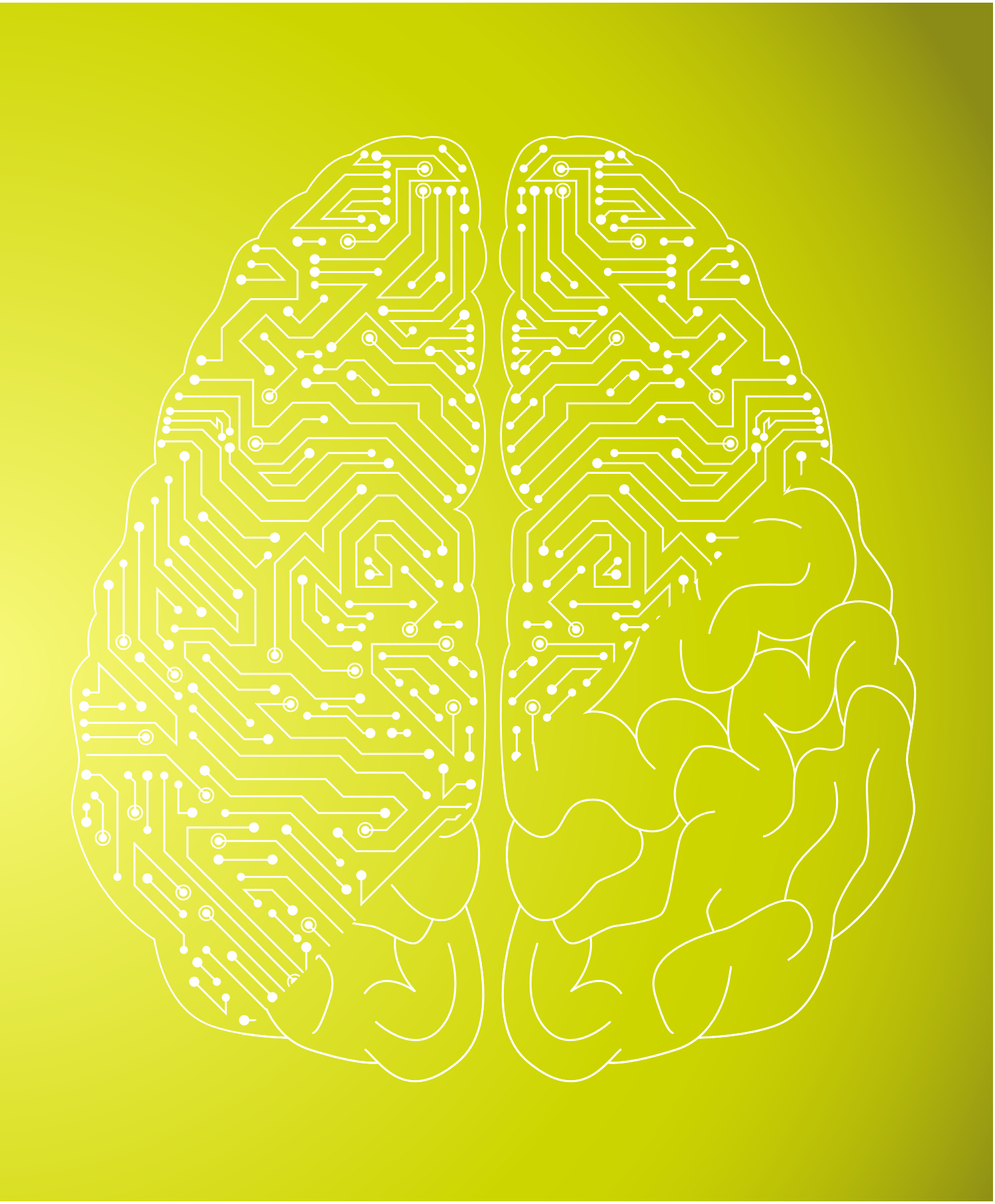
ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

A YEAR IN REVIEW 2021



Addressing global challenges in assuring the safety of robotics and autonomous systems

Key

- Funders
- Demonstrator projects
- Collaborative links
- Programme Fellows

Contents

Welcome

In parallel with technological advances, an intriguing story about the current landscape for introducing autonomy unfolded in the media in 2021.

We have seen robotic systems that aim to improve agricultural safety and tackle climate change, continuing coverage of AI in healthcare, reports of accidents involving autonomous vehicle systems, and drone delivery trials.

This increased visibility of fledgling autonomous systems impacts people's perception of what the technology can do. Owners of purportedly self-driving cars undertaking other activities while their car is in motion is a worrying example of this.

The results of public focus groups that we ran this year substantiate this view: that public perception of autonomous systems doesn't always match reality. This demonstrates wholeheartedly that we, as developers, assurers and regulators, need to be aware of public perception and must consider this as we develop, regulate, and market novel technologies.

2021 has been our busiest year yet. We strengthened our position as a provider of independent, practical guidance on the safety assurance of autonomous systems, with the launch of our methodology for the Assurance of Machine Learning for use in Autonomous Systems (AMLAS). This has been downloaded around 700 times by colleagues in numerous sectors in 18 countries.

We have again grown our team in York, including expertise in software safety assurance, assistive robotics, continuous assurance and cobots, as well as welcoming a new project manager to support our demonstrators.

Our work with regulators has developed, and their interest in our work and the guidance and expertise we offer has resulted in such organisations being a core part of many of our new demonstrator projects. While creating change in standards and regulations will take time, we are already ensuring that our guidance is considered by those regulating the incoming technologies.

We have also bolstered our work in education and training through bespoke courses, educational webinars and our own Advanced Topics in Safety MSc module. This helps to ensure that the people who are or will be developing and regulating autonomous systems are equipped with the skills and knowledge they need to put safety at the forefront of their work.

The AAIP team and community has come together again this year despite ongoing impact from COVID-19. I'm proud of them all and of what we have achieved. I look forward to more in 2022.

Professor John McDermid OBE FREng
Programme Director

Safety is a global issue. At Lloyd's Register Foundation we work with partners across the world to address the most pressing challenges we face as a society.

Assuring the safety of digital systems is a central focus of our strategy. Through our partnership with the University of York, the AAIP is developing guidance that can be used in any industry in any country to support the safe development and regulation of autonomous systems.

We must also consider the knowledge of those involved with these novel technologies. Through formal education, online tutorials, and bespoke CPD the AAIP is strengthening skills for safety in numerous sectors and countries.

Our aim as a charity is to engineer a safer world. Our work with the AAIP team is helping us to achieve this through their work to distill their peer-reviewed, expert research and evidence from demonstrator projects into guidance and training that is accessible to all.

Professor Richard Clegg FREng
Chief Executive
Lloyd's Register Foundation

Expanding and sharing knowledge is part of our vision to be a university for public good. For us, this is about collaboration: across disciplines and across geographical boundaries.

This principle is at the heart of the AAIP. From the beginning the Programme has funded projects that bring together developers, academics, regulators and professional bodies to establish the evidence needed to assure the safety of autonomous systems.

Bringing together philosophers, lawyers, scientists, engineers, clinicians and mathematicians is central to the Programme's research. Their interdisciplinary work welcomes those who are just starting out in their careers with others who are well established. These different backgrounds, expertise and experiences lead to a true sharing and expansion of knowledge.

Innovation and impactful progress are evident from this global, interdisciplinary approach. We know that assuring the safety of autonomous systems is a huge challenge. Through these collaborations, the Programme is advancing the way that we do this.

Professor Kiran Trehan
Pro-Vice-Chancellor for Partnerships and Engagement, University of York

A Year in Numbers

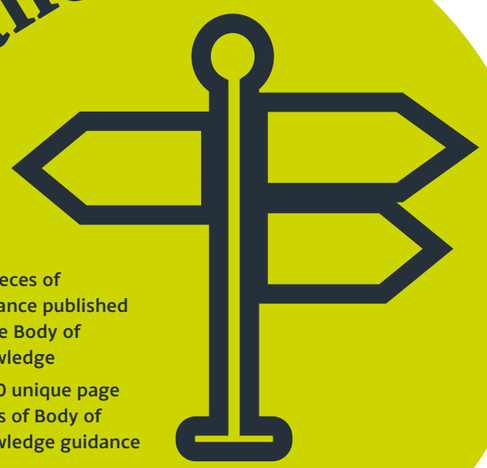
Research

18

active demonstrators

- 24% York team
- 76% demonstrators

Guidance



- 35 pieces of guidance published in the Body of Knowledge
- 5,600 unique page views of Body of Knowledge guidance

Funding leveraged

£39M

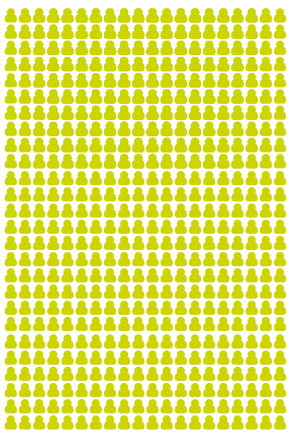
One new facility

Institute for Safe Autonomy built, with AAIP as founding partner

International community

Worked across **5** continents

Education and training



500+ people attended AMLAS workshops

50

health professionals on bespoke CPD

- UK – 8
- Germany – 4
- Belgium – 1
- Australia – 2
- UAE – 1
- Brazil – 1
- Iran – 1
- USA – 1
- France – 1

Dr Jordan Hamilton visits PAL Robotics in Barcelona, October 2021

International Community

Providing pragmatic, evidenced guidance, processes and techniques to those developing, researching and regulating autonomous systems has been central to our work across the globe.

We launched our methodology for the Assurance of Machine Learning for use in Autonomous Systems (AMLAS) in February. It has since been downloaded almost 700 times by developers, researchers and regulators in diverse sectors in 18 different countries. It is the first of the guides we will produce based on our research in key areas – machine learning, understanding, decision-making, societal acceptability and autonomous systems in complex environments (find out more on page 16).

This year we were delighted to host the 40th International Conference on Computer Safety, Reliability and Security (SafeComp 2021). Hosted as an online conference, we welcomed more than 100 international delegates to four days of workshops, keynote speeches, papers and social activities, with a special theme of safe human-robot interaction.

UK

We supported the development of a white paper, "Human factors and ergonomics in healthcare AI", published in September by the Chartered Institute of Ergonomics and Human Factors (CIEHF). The paper represents the work carried out by Dr Mark Sujan on an AAIP demonstrator project and was written in collaboration with colleagues from partner organisations.

We welcomed two new Fellows to the team. Dr Tom Lawton, a consultant and Head of Clinical AI at Bradford Teaching Hospitals NHS Foundation Trust, has published a number of papers with the team on the use of AI in intensive care situations including sepsis treatment and weaning patients off ventilators. Simon Smith is a Chief Architect and has been working with the AAIP team to consider our impact and how to extend the reach of our research.

Europe

We have three active demonstrators in Europe across different domains – healthcare, maritime and aviation. In October we visited our collaborators at PAL Robotics to discuss our joint project Ambient Assisted Living for Long-term Monitoring and Interaction (ALMI). The team in Barcelona showed us the test environment they have developed for the project, which is focused on the development of adaptation methods to enable assistive-care robots to cope with uncertainty and disruptions in a home environment.

We are pleased to be part of a partnership between AAIP, Fraunhofer IESE and Fraunhofer IKS. Layers of Protection Architecture for Autonomous Systems (LOPAAS) addresses core challenges in the safety assurance of autonomous systems and automated driving, in particular with regard to dynamic risk management and assurance of machine learnt (ML) components.

Asia

We were delighted to welcome a new Programme Fellow to the AAIP community. Mehran Alidoost Nia is a PhD student in software engineering at the University of Tehran. He's working with Professor Radu Calinescu on the verification of robotics and autonomous systems (RAS) via formal approximation techniques to support the verification of RAS at runtime.

Australasia

Our demonstrator with the Australian National University concluded this year, with an initial safety indicator developed and tested for autonomous driving scenarios. The team is now testing the mechanism on more complex scenarios.

We have continued our collaboration with the Trusted Autonomous Systems Defence Cooperative Research Centre in Australia. Together we have developed and run two industry webinars. The first was a general introduction to autonomous systems and safety assurance. The second was a focused look at our AMLAS methodology and how it can be used alongside the ML development process.

North America

We welcomed Dolline Hatchett, Director at the Office of Safety Recommendations and Communications for the National Transportation Safety Board in Washington DC to the Programme's Governing Council.

In August, we hosted an AMLAS tutorial at the International System Safety Conference, welcoming over 70 colleagues to a session on using the methodology.

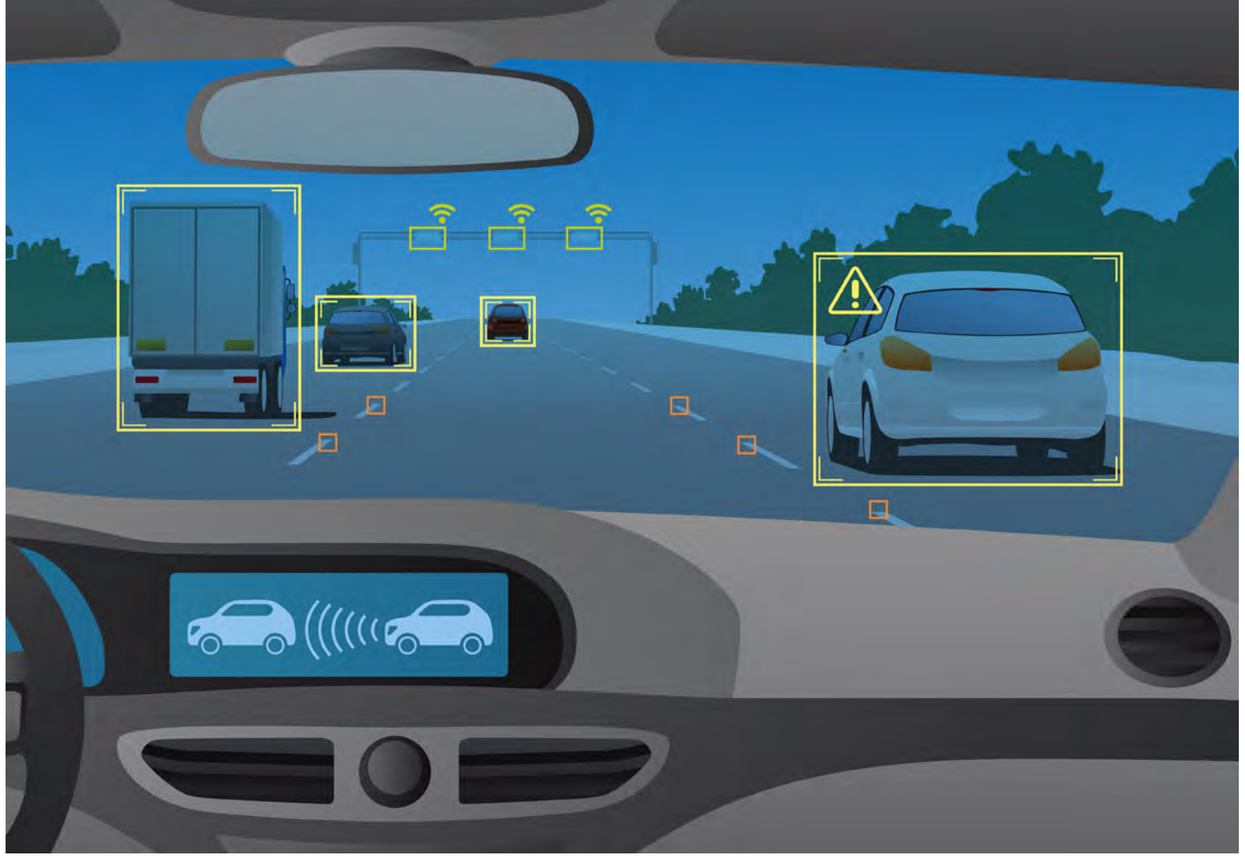
South America

Programme Fellow Dr Genaina Rodrigues and her research group have developed a framework for the modelling and decomposition of multi-robot system (MRS) missions. The framework supports the planning of MRS missions with interdependent tasks. It uses a goal-oriented modelling approach intertwined with a language for hierarchical MRS mission specification, and a systematic and automated mission decomposition process. Genaina and colleagues are also developing an architecture for mission coordination within heterogeneous MRS operating in disruptive environments and with constrained computational resources, and a simulation environment for the synthesis, execution and analysis of MRS missions.

Dr Richard Hawkins teaching on bespoke CPD with NHS Digital, November 2021

Public Engagement

We have become more reliant on digital technologies in the last 18 months. They have helped us to connect with family and friends, bring online events to life, and access shops and services during COVID-19 restrictions. How has this acceleration in technology use impacted our perception of autonomous systems?



In 2020 we ran focus groups to ask the public about autonomous technologies: about their perceived risks and their use in driving and healthcare. This year we wanted to ask these questions again and to see the impact on attitudes and perceptions, if any, of COVID-19 and the associated increased general use of digital technology.

Attitudes towards technology

As we saw in 2020, the groups we met spoke positively about how technology supports their lives. These benefits tended to be at an individual level: saving them time, helping them to keep fit, and to keep in touch with friends. They also had concerns and these were on a more societal level: that children might be missing out by being online, that vulnerable groups could be left behind, and a worry about technology replacing jobs.

There were two interesting new findings. Firstly, a generational difference in terms of technology fatigue: younger groups we spoke with were tired of the reliance on technology while older participants saw technology as novel and exciting. Secondly, a concern about the pace of technology and its environmental impact.

> *"...there is an awful amount of waste built into the sort of general leaping of technology. ...it is just literally more landfill. I think that's going to be a huge issue because obviously a lot of the components of the technology are toxic things which need processing."*

The societal impact arose again linked to the use of personal data. Participants were very unhappy about companies using their data, unless it was for the greater good.

Autonomous vehicles

Participants split naturally into three categories in their discussions of autonomous vehicles, as they had in 2020: supporters, potentials and rejectors.

Supporters believe that an autonomous vehicle would have undergone rigorous testing and that technology can outperform a human driver, similar to our results in 2020.

> *"The thing is, ... a computer can perform calculations and make decisions thousands of times faster than the human brain, so theoretically it should be safe."*

As part of our discussions, participants were shown a magazine cover that portrayed an autonomous car, and a short video showing a car identifying hazards. The film showing the hazard identification overwhelmed many of our participants, leading to a feeling of unease.

> *"Yeah, it was making me feel very uneasy thinking – seeing all those things. I think, if I reflect on it, if it was an autonomous system which had gone through all the safety precautions, then I would rather all the workings were just hidden for me."*

"Hiding the workings" is difficult, however, while there is a need for a human driver to be alert in case of a handover situation. Participants assumed that buying an autonomous car would mean giving full control to the system and many rejected the whole concept of a driverless car if they still had to be alert in case of an emergency.

> *"But I mean it kind of defeats the point, if I've got to sit in the front and be still very aware, you know, if my car is not necessarily going to take action for me and it's going to, you know, flag up warnings here and there and I've got to be ready, ...I might as well just drive the car normally."*

The feeling of apprehension caused by the video of hazard identification and the idea of handover led to conversations about what information they would want in such a situation. Participants were clear that they did not want to be over-stimulated by warnings and information. They wanted to know why they needed to take over the wheel and what was expected of them.

Healthcare

In contrast to autonomous driving, participants were clear that they wanted a human involved in our healthcare scenario, which involved patients monitoring their own blood pressure at home.

> *"I think as long as it complements the doctor's knowledge and not takes it over."*

> *"...the blood pressure monitor definitely seems like a good idea, as long as the machine is not making the decision."*

The groups identified benefits of the use of autonomous technology in healthcare, as long as it was for monitoring and not for taking the final decision about a diagnosis or a treatment.

Again our groups split into three categories: supporters, potentials and rejectors. Supporters were clear that technology could help to ease the burden on the NHS and could outperform a human. Potentials were uncertain and wanted more information about how easy such equipment might be to use, especially for older and more vulnerable groups. Rejectors were concerned about technology going wrong and most clearly wanted a human involved in the final decision making.

> *"Because in the medicine world there is not really room for mistakes, so I wouldn't say 100% start relying on machines."*

Next steps

We will undertake more in-depth analysis of the results of the focus groups to feed into our societal acceptance of autonomous systems research pillar. In addition we will host a number of events in 2022 that offer opportunities for members of the public to find out more about autonomous technology.

Education and Training

Our AMLAS methodology has offered an opportunity to reach a global audience with webinars and tutorials on the assurance of machine learning.

Nikita Johnson led safety training as part of the UK Manufacturing Robotics Challenge, July 2021

Over the year we met hundreds of safety engineers, clinicians, developers and others in a range of sectors. Through bespoke CPD, conference tutorials and webinars we have informed stakeholders about our process for assuring the safety of machine learning (ML).

We also strengthened our relationship with a number of partners to extend our training and education reach, and worked with young professionals in a global robotics challenge.

Academic education

We welcomed the first attendees to our new MSc module in April. The first course of its kind, it broadens students' existing system safety engineering knowledge to introduce the challenges autonomy and AI present to safety processes and product safety, and how to start addressing them. Our online offering included both pre-recorded and live streamed lectures and case study sessions. The Advanced Topics in Safety module will next run in April 2022.

Global reach

There has been wide reaching interest in our AMLAS methodology. It offers the first clear and detailed assurance process for ML components. It is complementary to the ML development process, which many stakeholders have found useful.

We were invited to run numerous bespoke, company or industry-specific webinars over the year, including for organisations from the German automotive industry and Dstl. A number of courses are planned with regulators and other organisations in 2022. We also hosted a three-hour tutorial for more than 70 delegates at the 2021 International System Safety Conference and were invited to give a talk to 100 safety experts at the 9th Scandinavian Conference on System and Software Safety in November.

We further developed our partnership with the Trusted Autonomous Systems Defence Cooperative Research Centre, with a focus on training and education for colleagues in Australia. Our first webinar focused on an introduction to ML and autonomy and the safety challenges that arise from these. Our second offered a focus on the assurance of ML. We will continue this partnership in 2022, looking at how we can support skills and knowledge development, particularly in the Australian maritime sector.

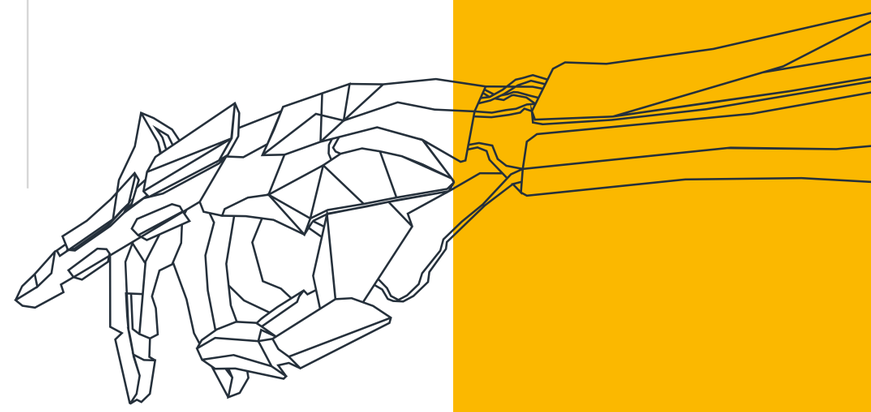
The next generation

We worked with more young safety professionals in 2021 through a partnership with the Advanced Manufacturing Research Centre, the UK-RAS Network, and Sheffield Robotics.

The Manufacturing Robotics Challenge took place online, bringing together around 40 participants to take part in the hackathon-style event on the safety and security of human-robot collaboration. We defined the safety requirements and safety assurance goals, gave background talks on what was required of the participants, and judged and gave feedback on the safety justifications that the teams produced as part of the challenge.

Bespoke CPD

Our partnership with NHS Digital continued in 2021. Together we ran two one-day Assurance of AI in Healthcare CPD courses for UK clinical risk managers and those working in health IT. We also joined forces with the Faculty of Clinical Informatics to host a half-day conference on the safe utilisation of AI in healthcare. NHS Digital's Clinical Director for Patient Safety, Dr Manpreet Pujara, welcomed more than 80 colleagues to the event, where we discussed defining regulation and safety strategy, robust assurance methodologies, clinical perceptions and trust in AI, and what we can learn from other industries.



"The MSc module...offers a well organised and structured introduction to the challenges to safety engineering raised by the introduction of robotics and autonomous systems. It added to my existing knowledge and the focused examples given were extremely useful. The team is helpful and patient and provide a friendly environment and great interaction with other students and teachers. I have already recommended the course to others."

Lorenzo Maldini, PhD student, University of Southampton

Let us develop a bespoke training course for your organisation: **bit.ly/aaipeducation**

Guidance

The Programme informs the safe development and regulation of autonomous systems by providing guidance that is freely accessible to all stakeholders in the community. This helps industry to generate the evidence needed to prove that autonomous systems are safe, and supports regulators in setting consistent safety standards worldwide.

> *"We are developing an AI flight controller for drones and safety assurance is a key aspect of our product roadmap. Without a way to assure the safety of our neural networks, there is no way to certify them and bring them to market. Despite the interest from government and industry, there is currently no established process for certifying and assuring an AI component. AMLAS fills this gap by providing us with a framework to integrate safety assurance into our development process and build a compelling argument for our safety case. We are very much looking forward to trialling the AMLAS process in our upcoming flight controller development project."*
>
> **Dr Matthew Carr, Co-Founder and Chief Executive Officer, Luffy AI**

To provide this expert guidance, the research undertaken as part of the Programme is translated into practical methods and processes to give confidence in the safety of the autonomous system. In 2021 we advanced this area of work when we published our methodology for the Assurance of Machine Learning for use in Autonomous Systems (AMLAS).

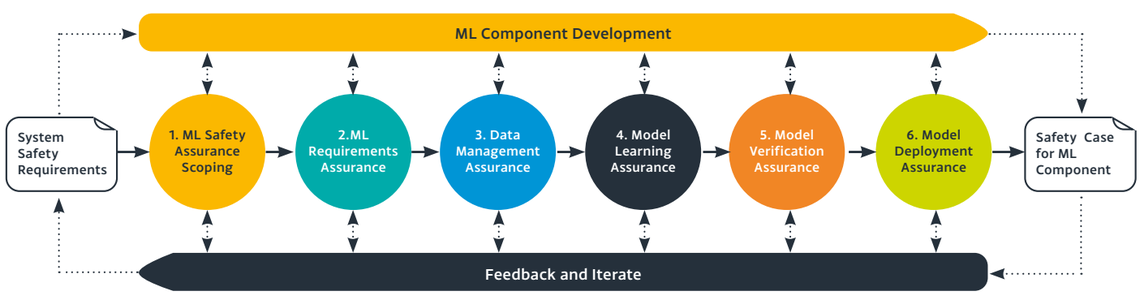
Core technical issues

The AMLAS methodology is the first guidance to be published from our research strategy comprising five key pillars (see page 16). It has reached across the world, with downloads in 18 countries across five continents and there has been considerable feedback about its applicability and usefulness.

Cross-domain guidance

Translating our foundational research into guidance is just one way in which we support the community. We also fund demonstrators in numerous sectors across the world, enabling teams to investigate how to generate the evidence needed to give confidence in the safety of an autonomous system. Over the year we published a number of pieces of guidance in our Body of Knowledge where these teams have turned their research into accessible overviews of their work, with recommendations or instructions on particular techniques or methods. While each project is focused on one particular domain, the guidance is generalised where possible to ensure it is usable by all.

We now have around 40 pieces of guidance across four fundamental areas of safety assurance in the Body of Knowledge: defining the required behaviour, implementing an autonomous system to provide the required behaviour, understanding and controlling deviations from this behaviour, and gaining approval.



Overview of the AMLAS process

The wider environment

The landscape for autonomous systems is complex and interconnected, regardless of the domain. Consequently, it is important for us to undertake additional research and write further guidance to give an understanding of the context in which autonomous systems will be deployed.

Accordingly, in 2021, funded by the Royal Academy of Engineering and Lloyd's Register Foundation's Safer Complex Systems programme, Professor John McDermid worked with three collaborators to develop a framework for understanding and improving the safety of complex systems.

Safer Complex Systems, An Initial Framework, July 2021

The report, 'Safer complex systems: an initial framework' was published in July 2021. The framework has since been applied in work undertaken in partnership with Egis for UK Research and Innovation to gain insight into the potential considerations for the safety of emerging complex systems in future flight. This analysis supported the development of the Future Flight Aviation Safety Framework.

We have also been involved in standards work around the use of data in autonomous vehicles. Dr Mark Nicholson was part of the steering group that developed PAS 1882:2021 – Data collection and management for automated vehicle trials for the purpose of incident investigation. This is the first consensus standard to enable data collection and management for automated vehicle trials to support incident investigation. It specifies requirements for the collection, curation, storage and sharing of information during trials in the UK in relation to information collected or received by the car.

Utilise our guidance:
bit.ly/aaipguidance

Research Strategy

To genuinely impact robotics and autonomous systems (RAS) and ensure they benefit society as a whole, our work must be based on both sound academic research and empirical study. This ensures that our research is validated "in the wild" and provides evidence that is applicable to those developing and regulating the technology.

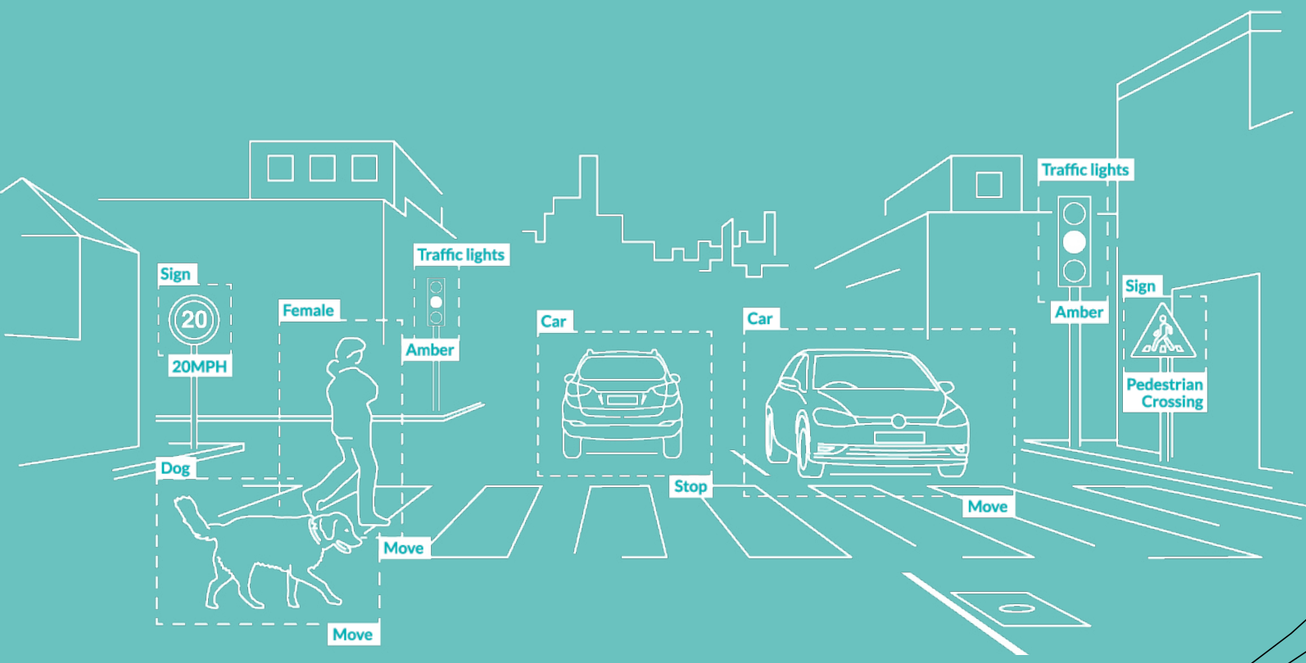
Our approach distinguishes us from others in the field. It is at the heart of the strategy we refined over the last year, which is based on three guiding principles that define a journey from academic research to useable evidence:

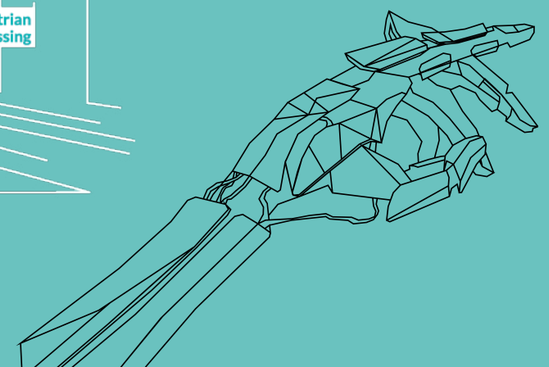
1. Sound research – peer-reviewed and published in leading academic venues
2. Empirical evaluation – evaluated in credible and real-world contexts
3. Practical and accessible – disseminated online and through CPD training

Our collaborative, evidence-based practice is reflected in AMLAS, our methodology for the Assurance of Machine Learning in Autonomous Systems. This was initially developed by the York team, building on their earlier published work on assurance argument patterns and emerging evidence from demonstrators. It was refined through our work with Programme Fellows and then peer-reviewed by Fellows and other key stakeholders. It was then published and disseminated through tutorials, bespoke CPD, webinars, workshops and conferences.

The strategy is based on core technical issues that must be considered for the safe development and introduction of any autonomous system. It has five pillars:

1. Assurance of Machine Learning for use in Autonomous Systems (AMLAS). AMLAS is the first published guidance from the strategy.
2. Safety Assurance of Autonomous Systems in Complex Environments (SACE)
3. Safety Assurance of Understanding in Autonomous Systems (SAUS)
4. Safety Assurance of Decision-making in Autonomous systems (SADA)
5. SOCial Acceptability of autonomous systems (SOCA)





AMLAS

This is the first assurance process of its kind. It provides a clear and detailed methodology for machine learning (ML) components used in autonomous systems. It is complementary to the ML development process, sitting aside the key stages in the ML lifecycle. It enables the generation of the evidence needed for explicitly justifying the acceptable safety of these components when integrated into an autonomous system.

AMLAS comprises a set of safety case patterns and a process for systematically integrating safety assurance into the development of ML components. This provides a compelling argument about an ML model to feed into a system safety case.

SACE

SACE will provide similar guidance to AMLAS but at a system level. It covers:

- the elicitation of safe autonomous system behaviour in complex environments
- analysing interactions between autonomous systems and the outside world, including humans
- the validation of safe autonomous system behaviour in complex environments, including the use of simulation
- maintaining safety assurance of an autonomous systems during operation
- creating a safety case for autonomous systems

It is being peer reviewed by Programme Fellows and other experts from a range of domains and backgrounds and will be published in 2022.

SAUS

This pillar is focused on the understanding element of an autonomous system. It will comprise:

- the elicitation and validation of safety requirements for understanding (e.g. perception) in autonomous systems
- failure analysis and propagation for understanding
- verification of understanding (e.g. perception)
- creating safety case for understanding in autonomous systems

SADA

The decision-making elements of the autonomous system are the focus of the SADA pillar, which will incorporate:

- the elicitation and validation of safety requirements for decision making (e.g. path planning) in autonomous systems
- failure analysis and propagation for decision making
- verification of decision making (e.g. path planning)
- creating a safety case for decision making in autonomous systems

SOCA

Moral and legal governance of autonomous systems, alongside societal acceptability, run through each of our research pillars but are the sole focus in the SOCA pillar. Specifically, this pillar will consider:

- legal acceptance
- regulatory compliance
- accounting for ethical considerations
- risk acceptance
- public trust

Independent but interconnected

Each pillar stands alone, with the guidance that emerges from each able to be used to support the safety assurance of a specific component within an autonomous system. They are also interconnected, as the components within a system are. Used together the guidance documents from each pillar will help to ensure a credible and compelling assurance case is created for an autonomous system.

View the research strategy **bit.ly/aaipresearchpillars**

Towards safe robotics and autonomous systems: how our research strategy translates into impact

Assurance of Machine Learning in Autonomous Systems (AMLAS)

Lead researchers: Dr Richard Hawkins and Dr Ibrahim Habli

AMLAS is already used by colleagues across the globe to support their work to assure the safety of their machine learnt (ML) components. In the last year we have been further evaluating the methodology through the development of case studies in various domains, including healthcare, space, manufacturing and automotive.

We have disseminated AMLAS through numerous channels including workshops, bespoke CPD, our MSc modules (Computers and Safety and Advanced Topics in Safety) and numerous webinars and tutorials. AMLAS is a core part of four demonstrator projects: AI Clinician, SAFR, ASSIST, and ACTIONS (see pages 22 to 33). These projects are either trialling AMLAS for specific technologies or adapting the guidance for a particular domain.

Significant progress has been made on a tool that will support the use of AMLAS. This will help users with the AMLAS workflow and creating the required artefacts and will enable integration with existing engineering models and tools, and the automation of the instantiation of the safety argument. This will be launched in 2022 along with an interactive version of the guidance accessible through a standalone website. In 2022 we will also start work on a new multidisciplinary, cross-institutional project that will include further development of AMLAS, in particular to consider through-life learning and broader ethical and legal considerations.

> *Working in AAIP gives me the opportunity to discuss open research questions about the safety of machine learning when used in automated vehicles with experts and to look at them from different perspectives. The knowledge from other fields can help to highlight new approaches, requirements and contexts. AMLAS has made an important contribution to safety case patterns and provided valuable guidance on how to proceed to justify the safety of machine learning components."*
>
> **Lydia Gauerhof, Research Engineer, Robert Bosch GmbH**

Publications

- Download the AMLAS methodology: bit.ly/aaipamlas

Safety Assurance of autonomous systems in Complex Environments (SACE)

Lead researchers: Dr Richard Hawkins and Mike Parsons

The focus of SACE is on the overall system-level assurance activities, particularly considering the interactions of the autonomous system with its complex environment. Similar to AMLAS, this guidance will define a safety assurance process along with corresponding safety case patterns.

The stages of the methodology have been identified:

1. Operating context assurance
2. Identification of autonomous system hazardous events
3. Safe operating concept assurance
4. Autonomous system safety requirements assurance
5. Autonomous system design assurance
6. Management of hazardous failures
7. Out of context operation assurance
8. Autonomous system verification and validation

A wide range of examples are being developed to support the guidance material. The guidance has been informed by the safety assurance work being undertaken as part of the development of a team of robots that will be used in the new Institute for Safe Autonomy (ISA) facility.

It is written in a style and format that is accessible to regulators and engineers. Considerable progress has been made on the first draft and it will be ready for review in the first quarter of 2022.

> *Safety assurance of autonomous systems requires the consideration of the whole system in its broadest context. This includes considering the environment in which the autonomous system operates, as well as the interactions that the system has with other things in that environment. We can't reason about the safety of any part of an autonomous system, such as its sensors or its machine learning components, without considering their impact on the overall system. SACE is developing a safety assurance process that enables the development of a safety case for the autonomous system as a whole. This provides a framework within which more specialised assurance aspects, such as AMLAS, can be integrated."*
>
> **Dr Richard Hawkins, Senior Research Fellow, AAIP**

Safety Assurance of Understanding in autonomous Systems (SAUS)

Lead researchers: Professor John McDermid and Dr John Molloy

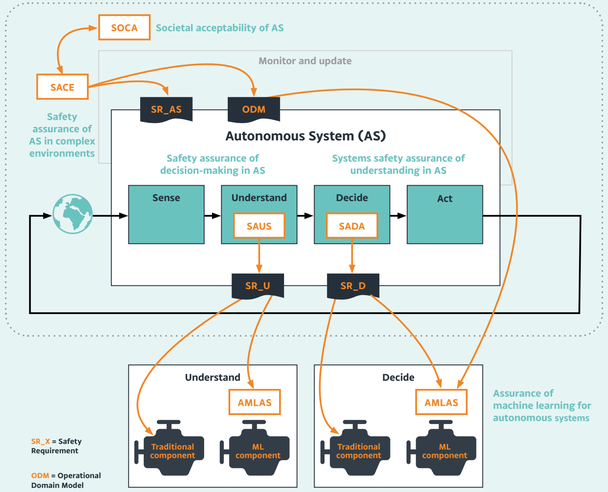
The focus for SAUS is split into two principal areas of research. First is work to identify and explain the nature of failures in the understanding component of autonomous systems. This work examines failures that exist owing to the technical and theoretical limitations of sensors, as well as those that can arise because of limitations in the ML algorithms that are used to build a model of the system's environment (e.g. to classify objects). These failure modes have been identified through observation and analysis of deployed systems, particularly in the automotive domain.

The second principal area is developing an adaptation of the Hazard and Operability Study (HAZOP) process to the safety assurance of the understanding component of systems. This is informed by the theoretical and observed failure modes identified in the first area of research. Current adaptations of HAZOP to computer-based systems use relatively simple guidewords to prompt identification of potential hazardous failure modes. SAUS is developing a much richer set of guidewords, closer to those used in the initial definitions of HAZOP in the process industries, to deal with the fact that autonomous systems process much more complex sensor data (e.g. video, and lidar point clouds) than earlier computer-based systems. Confidence in this approach is found in the observation of the hypothecated failure modes in reports relating to deployed systems (e.g. fleets of road vehicles) as they become more widespread.

This work is closely aligned with the safety assurance work on the team of robots that will be deployed in the ISA building and will use the robots' perception suites as a case study.

> *RAS build a model of the world through their sensors and the understanding component. Errors in this world model (e.g. mislocating the RAS, or misclassification of objects), can contribute to or even cause accidents, as has been seen in several case (e.g. the Uber Tempe accident). Thus, assessing the potential failure modes and behaviour of the understanding component of a RAS is critical to designing a system to minimise errors, and to overall safety assurance."*
>
> **Dr John Molloy, Research Associate, AAIP**

In lieu of standards and regulations for robotics and autonomous systems (RAS), we need new safety processes and methodologies to generate the evidence needed to prove autonomous systems are safe. Translating the research undertaken as part of our programme into free to access, expert guidance bridges this gap while regulations are established. This is our approach to impacting the safety of RAS.



AAIP's five research blocks and their relationships to the autonomous system architecture

SOCial Acceptability of autonomous systems (SOCA)

Lead researchers: Dr Ibrahim Habli and Dr Zoë Porter

Research in this pillar has followed two main strands. First, the development of a methodology for including ethical principles in the development and assurance of autonomous systems. A workshop held in January 2021 helped to shape the direction of this work. The event, 'From Ethical Principles to the Ethically-Informed Engineering of Autonomous Systems', brought together members from the engineering, ethics and regulation communities. Second, the expansion of our interdisciplinary research on responsibility. This has led to an ambitious project, funded by the UKRI's Trustworthy Autonomous Systems Programme and commencing in January 2022, to develop a framework for tracing and allocating responsibility for autonomous systems. This project brings together engineers, lawyers, philosophers, developers and the public to address one of the most important unanswered questions about autonomous systems, and a condition of their societal acceptability: who is responsible for the decisions and outcomes of autonomous systems?

We continued to work with regulators and policymakers on the creation of an effective governance ecosystem for autonomous systems, including close engagement with the Law Commission on anticipatory regulation for autonomous vehicles.

> *Autonomous systems are being developed for complex, real-world environments where they may affect not just the physical safety of people but also their psychological well-being and fundamental rights. The deployment of these systems also raises broader societal questions about equitable distributions of risk and benefit, environmental sustainability, and the allocation of responsibility. It is crucial that we address this complex range of issues effectively if we are to achieve trust in autonomous systems and harness the benefits that they can bring."*
>
> **Dr Zoe Porter, Research Associate, AAIP**

Publications

- Porter, Z. "From Ethical Principles to the Ethically-Informed Engineering of Autonomous Systems: workshop report", February 2021 (published on the AAIP website).
- McDermid, J., Jia, Y., Porter, Z., and Habli, I. "Artificial intelligence explainability: the technical and ethical dimensions" in Philosophical Transactions of the Royal Society A, August 2021.

Safety Assurance of Decison-making in Autonomous systems (SADA)

Lead researcher: Professor Radu Calinescu

Two main research strands have been explored as part of this pillar. As part of the first strand a Decision-Point Analysis (DPA) process was developed for the analysis and assurance of autonomous decision-making. The new process builds on existing safety engineering techniques. It involves the creation of activity diagrams for the autonomous decision-making workflow, performing DPA and targeted HAZOP on these diagrams, for the purpose of eliciting safety requirements.

The second research strand has focused on autonomous systems that use deep neural network classifiers for the perception step of their decision-making. The work is developing a new method for the correct-by-construction synthesis of discrete-event controllers for these systems. Progress has been made with the development of case studies from the mobile robot navigation domain (focusing on collision mitigation) and shared-control between human drivers and Level 3 automated driving systems (focusing on managing the attentiveness level of the human driver).

The research in this pillar is closely related to a number of AAIP demonstrators with tasks that address challenges of autonomous decision-making, including ALMI, CSI:Cobot, SafeSCAD and SAFEMUV (see pages 22 to 33).

> *Autonomous systems are entrusted with decisions that have hitherto been a human prerogative. Increasingly, this happens in safety-critical applications from domains ranging from health and social care to transportation and manufacturing. The theory and practical tools developed by our SADA pillar support the assurance of autonomous decision-making by enabling the elicitation of the relevant safety requirements, the quantification of the uncertainty under which autonomous systems make decisions, and the principled engineering of the autonomous system components responsible for these decisions."*
>
> **Professor Radu Calinescu, Professor of Computer Science, University of York**

Publications

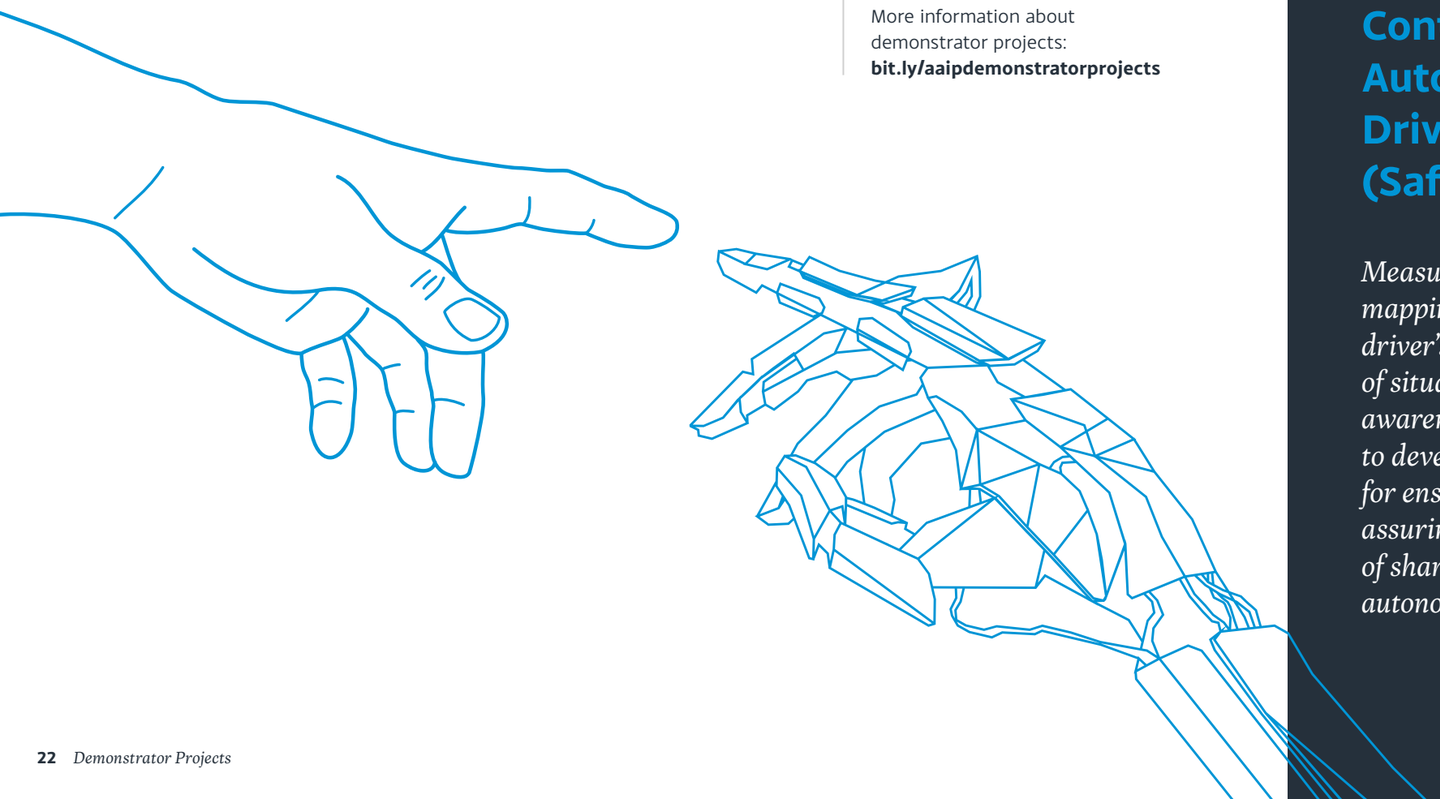
- Douthwaite, J.A., Lesage, B., Gleirscher, M., Calinescu, R., Aitken, J.M., Alexander, R. and Law, J., 2021. A Modular Digital Twinning Framework for Safety Assurance of Collaborative Robotics. Frontiers in Robotics and AI, 8, 2021.
- Gleirscher, M., Calinescu, R., and Woodcock, J. "RiskStructures: A Design Algebra for Risk-Aware Machines" in Formal Aspects of Computing, 2021.
- Gleirscher, M., Johnson, N., Karachristou, P., Calinescu, R., Law, J., and Clark, J. "Challenges in the Safety-Security Co-Assurance of Collaborative Industrial Robots", in The 21st Century Industrial Robot: When Tools Become Collaborators pp 191-214 (Springer).
- Gautam, V., Gheraibia, Y., Alexander, R. "Runtime Decision Making Under Uncertainty in Autonomous Vehicles", Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2021).

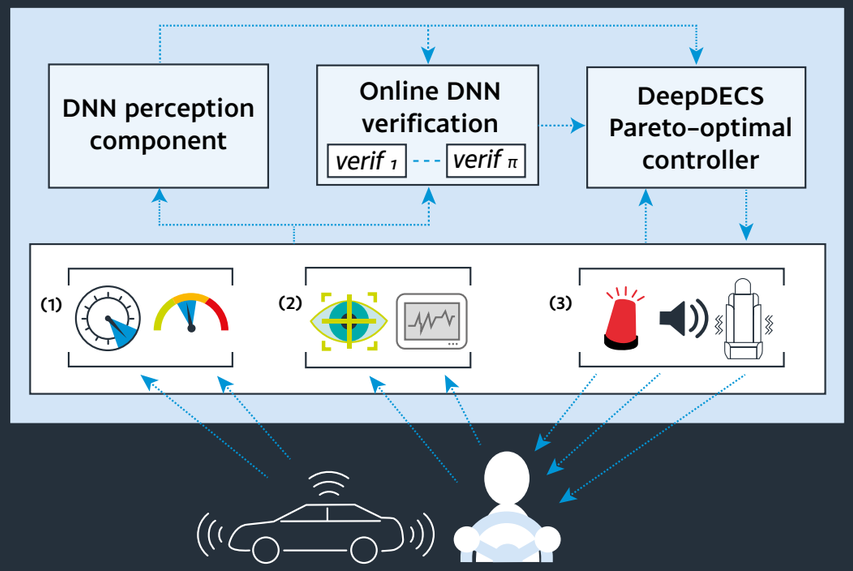
Demonstrator Projects

Our demonstrator projects contribute evidenced, repeatable techniques for demonstrating the safety of autonomous systems. We published multiple pieces of guidance in the Body of Knowledge over the year, written by the demonstrator teams as their projects reach completion.

The demonstrators are an important part of the international assuring autonomy community that has built up over the life of the programme. Their views on the landscape we're all working in offer practical insights. In this section you will find their responses to some of the questions others in the community might also be facing.

More information about demonstrator projects:
bit.ly/aaipdemonstratorprojects





SafeSCAD driver-attentiveness management system. Data from car sensors (1) and driver biometric sensors (2) are supplied to a DNN perception component that classifies the driver state as attentive, semi-attentive or inattentive. The DeepDECS controller decides when optical, acoustic and/or haptic alerts (3) should be used to increase the driver's attentiveness.

Safety of Shared Control in Autonomous Driving (Safe-SCAD)

Measuring and mapping a safety driver's level of situational awareness in order to develop methods for ensuring and assuring the safety of shared control in autonomous driving.

How can the verification of neural networks and of traditional software components be combined to provide assurance evidence for systems comprising both types of components?

Like many safety-critical autonomous systems, the driver-attentiveness management solution prototyped by SafeSCAD combines conventional software and deep learning components. Integrating the two paradigms is essential for ensuring that Level 3 autonomous car drivers retain sufficient situational awareness to safely take over the driving task when the car approaches traffic conditions outside its operational design domain. Driver biometrics and car parameters collected by specialised sensors are fed into a deep neural network (DNN) responsible for predicting the driver's response time to a potential control takeover request from the car. Based on the DNN predictions, a conventional software controller issues visual, acoustic and/or haptic alerts when the driver is deemed overly distracted.

As traditional verification methods cannot provide safety guarantees for systems mixing conventional software and deep-learning components, the use of DNN perception within our solution posed major challenges for its assurance. We overcame them using Deep-learning aware Discrete-Event Controller Synthesis (DeepDECS), a novel hybrid verification approach co-developed with colleagues from the SADA AAIP pillar (see page 20). DeepDECS uses a combination of DNN verification methods – both when the autonomous system is developed, and during its operation:

- At development time, DNN verification is used to quantify the aleatory uncertainty of the deep-learning perception, so that traditional verification methods can synthesise conventional software controllers aware of the DNN-perception uncertainty.
- In operation, online DNN verification associates trustworthiness levels with deep-learning predictions, enabling the controllers of autonomous systems to react confidently to trustworthy DNN outputs, and conservatively to DNN outputs that cannot be trusted.

Professor Radu Calinescu
Professor of Computer Science
University of York

Sense – Assess – eXplain (SAX)

Developing autonomous vehicles (AVs) that can sense and fully understand their environment and explain the decisions they take.

Bounding box of road user and related road commentary

How can explainability support the overall assurance of AVs?

It is critical that when AVs are deployed they are safe, accountable, and trustworthy. To be safe, AVs must identify, assess, and mitigate risks. However, to be accountable, they must do this in ways that users, developers, and regulators understand what the cars have seen, what they have done, what they are planning to do, and why.

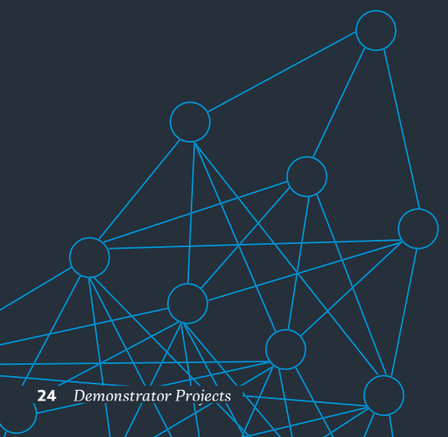
Let us consider the recent fatal crash of a self-driving car when it did not recognise a pedestrian (https://en.wikipedia.org/wiki/Death of Elaine Herzberg). Imagine what the car could explain to the driver before requesting assistance prior to the crash, what information developers would need when debugging causes, or what regulators would require when investigating the crash. Post-hoc explanations containing the vehicle's observations of other road users, traffic signs as well as road rules it acted on can serve as evidence to the causes of an accident and inform the investigation.

Regulators require some form of interpretability and explainability. For example, the European Union's General Data Protection Regulation (GDPR) and the European Parliament's resolution on "Civil Law Rules on Robotics" guarantee meaningful information about the logic involved in certain automated decisions[1]. Moreover, the GDPR advocates the "right to explanation" as a potential accountability mechanism, requiring certain automated decisions (of AI and robotic systems) to be explained to individuals.

In the SAX project we have designed, developed, and evaluated technologies that allow AVs to understand their environment, assess risks, and provide causal explanations for their own decisions. We conducted a field study in which we deployed a research vehicle in an urban environment. While collecting sensor data of the vehicle's surroundings, we also recorded an expert driver using a think-aloud methodology to verbalise their thoughts. We analysed the collected data to uncover the necessary requirements for effective explainability in intelligent vehicles. We show how intelligible natural language explanations that fulfil some of the key elicited requirements can be automatically generated based on observed driving data using an interpretable approach. These transparent and interpretable representations will enable developers to analyse an AV's behaviour and assure its safe autonomous operation. Users will also benefit from explanations by developing trust in autonomous vehicles.

Dr Lars Kunze
Departmental Lecturer in Robotics
Oxford Robotics Institute

1. S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, explainable, and accountable AI for robotics," Science Rob., vol. 2, no. 6, 2017.



Safety of the AI Clinician

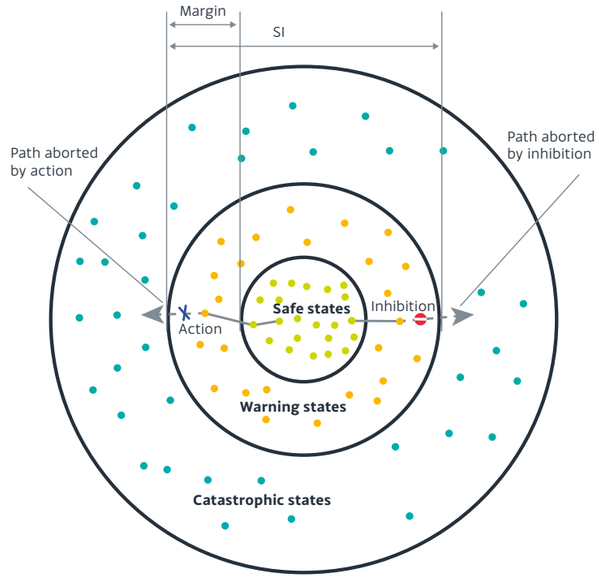
Investigating how to assure the safety of an AI-based decision support system (DSS) for sepsis treatment in intensive care.

How can the safety of AI and autonomous systems be assured in complex healthcare settings, especially in cases where there's a lack of consensus on accepted clinical practice?

Much of the research into AI-powered clinical decision support models focuses on assessing the performance or effectiveness of these technologies, with very little work on the systematic assessment of their safety (e.g. the identification, analysis, and elimination or control of clinical hazards throughout a system's life-cycle).

Establishing this basis for safety assurance is essential if AI decision support models are to be validated and converted into medical devices for use in clinical practice. Existing concepts for safety engineering can be used and have been formalised in safety assessment frameworks such as the Safety MOnitoring Framework for Autonomous Systems (SMOF).

In our work we applied the SMOF methodology to our AI Clinician model, an algorithm that informs the treatment of sepsis. A cornerstone of sepsis management includes the administration of intravenous fluids and/or vasopressors to restore a normal circulating blood volume and prevent further organ dysfunction. However, determining the correct dose and timing of these interventions, as well as resuscitation goals, is highly challenging for human doctors. Individual patient requirements vary substantially, and even in the same patient, treatment requirements can change rapidly.



Partition of system states in catastrophic, warning and safe states

Recreated from Machin, M, et al SMOF – A Safety MOnitoring Framework for Autonomous Systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems, IEEE, 2018

To circumvent the lack of consensus about what represents a safe and effective treatment strategy, we used domain expertise to define scenarios likely to represent unsafe decisions. An example of such a decision would be to not intervene on a patient with very low blood pressure, because it would precipitate the onset of organ failure.

With a set of scenarios defined we then assessed both human (clinicians whose data is observable in our training database) and AI agent behaviour against them. We demonstrated that the AI Clinician was 6% less likely than human clinicians to recommend actions that we labelled as unsafe. Further, we updated the model learning by reshaping the reward function, and penalised the agent during model learning when it made unsafe decisions. By doing this, we were able to further increase the safety of the AI agent: compared with human clinicians the new version of the model is 12% less likely to suggest decisions labelled as unsafe, without a significant drop in model performance.

To our knowledge, this work is the first successful attempt to define and test safety requirements for a clinical DSS based on reinforcement learning, considering multiple clinical hazards, and successfully modify the reward function of such an agent with added safety constraints.

These advances provide a use case for the systematic safety assurance of AI-based clinical systems, towards the generation of explicit safety evidence, which could be replicated for other AI applications or other clinical contexts, and inform medical device regulatory bodies.

Dr Matthieu Komorowski
Consultant in Intensive Care, Charing Cross Hospital and Clinical Senior Lecturer, Imperial College London

Ambient Assisted Living for Long-term Monitoring and Interaction (ALMI)

Demonstrating how novel robotic technology, environment monitoring capabilities, verification techniques, and adaptation methods can be integrated and applied to address concerns for autonomous robots used in people's homes.

Assistive care robots are required to work close to or alongside potentially frail humans. What changes to the robotic platform are required to mitigate the safety risks associated with this requirement?

We are reviewing the arm design of our TIAGo robot to make it safer for collaborative tasks with more vulnerable people. The redesign of the arm is focused on reducing the risks associated with the force that the robot could use in its tasks. The new prototype arm will have the following features:

- brakes in every joint to increase safety in case of an emergency stop
- every joint will have Series Elastic Elements (SEE) for torque sensing
- a new wrist design that includes three sequential joints to remove potential backlash
- using EtherCAT communication bus to increase software control bandwidth

This new version will pave the way for developing torque control at joint level and also in the operational space. The integration of torque sensors will allow the implementation of force control, enabling human-robot interaction to be performed with a much higher level of risk prevention, safety and effectiveness in activities where manipulation requires limiting the forces exerted by the robot arm. Ultimately our redesign will deliver a sensorized robotic arm with a safe control technique for advanced force control with online trajectory adaptation to increase safety in human-robot interaction.

Jordi Pagès
Head of Intra-logistics & Retail Solutions and TIAGo Product Manager
PAL Robotics

Human Factors in the design and use of Artificial Intelligence in healthcare (HF/AI)

Publishing guidance for regulatory bodies and technology developers on using HF in the design and use of AI in healthcare, through a white paper issued by the Chartered Institute of Ergonomics and Human Factors (CIEHF).

How does addressing human factors from a systems perspective help us assure the safety of AI in healthcare?

AI in healthcare is big news. Every day we see another story about a new app or system and we've seen some very encouraging results, particularly in diagnostics such as breast cancer screening. This is really positive, and this research could make a real difference.

We need to be aware that most examples of healthcare AI to date have been evaluated retrospectively. So, what we're really seeing are AI technologies that, in isolation and based on high-quality data, perform well.

The issue is when AI technology is used in a complex context such as a hospital. Focusing only on the technology during the development phase could lead to an unsafe situation when introducing it into the complex clinical setting. By using a human factors and ergonomics (HF/E) approach we take a systems perspective to technology development to help ensure the AI works as expected out of isolation and in the real world.

To support developers and other healthcare stakeholders, with colleagues I wrote a white paper published by the Chartered Institute of Ergonomics and Human Factors. The paper outlines eight HF/E principles to consider when designing an AI healthcare application to help assure its safety.

These considerations include well-understood principles from experiences with highly automated systems introduced from the 1970s onwards, such as workload, which already have methods and frameworks associated with them. Other principles covered in the white paper also exist with automated systems but are changing and becoming more complex because of the introduction of AI. For example, situation awareness: with the introduction of AI, both the human and the system need awareness, and technology developers must consider how the AI develops this awareness and communicates it to others.

There are also entirely new principles or ones that are more relevant because of AI. In particular, the relationship between staff and patients. When a nurse checks an infusion pump it's about much more than a technical adjustment. It's about checking in and finding out how the patient is really feeling, and patients and staff think this is really important.

At its heart, healthcare is a relationship between the patient and the clinical team: it's about humans. AI can support this, but the technology must be right, not just in isolation but also in the messy, complex system that is a hospital. Technology is one part of the story. To assure the safety of AI in healthcare, we must remember the human and a systems perspective introduced through HF/E is the way to do this.

Dr Mark Sujan
Director
Human Factors Everywhere

Download the white paper:
bit.ly/aaiphumanfactors

Assistive robots in healthcare

Investigated and evaluated the safety and regulatory requirements of close human-robot interaction in unstructured domestic environments, utilising the CHIRON robotic system.

What is distinctively challenging about training users (of all kinds) to ensure the safety of an assistive autonomous system?

The users of autonomous assistive systems, especially of those being designed to support independent living, will include health care professionals (HCPs), patients, relatives and a wider circle of informal carers. This presents a distinctive challenge when considering the training required by each group. Not only will there be different responsibilities when it comes to ensuring the operational safety of these systems, but the operational parameters of such systems will require adaptation to match the diverse and changing clinical needs of the patient. The training for HCP configuring these systems needs to encompass how these systems should be setup or adapted to suit the specific level of cognitive, sensory and physical impairments of the patient, together with the environment that they are used in, which is likely to be highly dynamic. In our research we have found that current assistive robotic systems, and the standards that define their design requirements, do not consider how such safety aspects relating to vulnerable users with accessibility needs (physical, sensory and cognitive) would be configured and who would need to take responsibility for ensuring that this is done appropriately – the autonomous system, or the HCP.

In a recent survey we conducted with HCPs we identified that in order to ensure safe operation, HCPs must know how to interrogate and evaluate the system, which might have autonomously adapted its behaviour or functionality in response to the changing needs of the patient. They would also need to assess whether the system was still safe from a clinical point of view, whether or not it was over- or under-supporting the patient. Over-support by a system, such as an assistive robot helping someone stand up, could result in the patient becoming weaker over time. Care providers also have to make sure that the equipment they provide is properly maintained, and as such it would fall on them to carry out routine inspections and verify system checks on sensors being clean and calibrated. The patients would need training on how to interact with the system to ensure that its behaviour was as they desired, otherwise they are at risk of losing their autonomy. As with any technology, there are likely to be times where only the patient's relatives or informal carers are available, and as such, would also require guidance on what they should or shouldn't adjust or change.

It is only by fully engaging with all users of these technologies to understand their training needs that we will see these systems accepted or adopted.

Professor Praminda Caleb-Solly
Professor of Embodied Intelligence
University of Nottingham

Workshop with healthcare professionals (photo credit: Hazel Boyd, Designability)

Safety Assurance FRamework for machine learning in the healthcare domain (SAFR)

Establishing a safety assurance framework to support healthcare manufacturers and deploying organisations to assure their ML-based healthcare technology and meet their regulatory requirements.

What are the current challenges for integrating the different technical, clinical and organisational perspectives needed for a whole system approach to safe AI in healthcare?

Safe and effective care delivery is a complex socio-technical system-of-systems challenge characterised by multifaceted care pathways, serving patients with unique and varied physiological conditions and delivered by an integrated team of care professionals with different skills, competencies, and experience. Safety management is very different from other domains, where the focus is maintaining a safe state to one where management focuses on moving from a state of high risk to one of lower risk. AI has the potential to improve both the safety of patients and the efficiency of the care service, however the much needed evidence to corroborate this is still being gathered due to the intrinsic nature of AI output being non-deterministic. This leads to challenges on understanding how its impact is measured from a safety perspective.

Whilst the concept of AI is familiar with care practitioners, there is a prevailing (mis)conception that it is roaming free, learning in real-time, and making decisions autonomously. This needs to be addressed and the near-term capabilities, limitations and opportunities understood. Invariably, care is delivered by practitioners who are integrated into and form the aggregating actuator in the pathway. AI technology must be integrated into a care pathway in such a way that practitioners do not lose their situational awareness and are able to intervene and take control of an escalating situation. Undoubtedly, care practitioners will need to learn new skills, but it is imperative that adoption of AI supported care services doesn't lead to skill fade in the fundamental science of care management and delivery.

The regulatory landscape across the care domain is complex and fragmented with a variety of different organisations and approaches involved. Technology manufacture falls within the scope of one of two regulatory regimes, based on its intended use and risk profile; in some circumstances it may fall in both. Whilst there are some areas of convergence in regulation there are also significant differences, but in essence access to market is granted through inspection of work processes and demonstration of the product's safety characteristics. The approach to deployment regulation is very different where it's the performance of an organisation delivering a care service that is appraised with little consideration given to the technology that supports the service. It is recognised across the health domain that existing regulations, standards and guidance are lacking in terms of supporting the safe development and use of AI systems.

Significant investment is being made to address this through initiatives such as public consultation on future regulation of medical devices; development of new standards through the newly established AI international standards committee, ISO/IEC JTC 1/SC 42, and the introduction of a Multi-Agency Advisory Service to signpost to resources.

Sean White
Senior Safety Engineer
NHS DIgital

ASsuring Safe artificial Intelligence in ambulance Service 999 Triaging (ASSIST)

Improving the recognition of out-of-hospital cardiac arrest by using an AI system to support ambulance service call centre staff.

How can an appropriate balance between human and system authority be determined, when using AI-based advisory systems?

Determining this balance can be challenging but is essential in healthcare settings due to the critical and unpredictable nature of the environment. There are many factors to examine and these must be considered continuously, before, during and after the implementation of the system.

Firstly, it is important to ensure that the system is always implemented as a support tool that works alongside healthcare professionals (HCPs) and not as a substitute for care, as although AI is quick and efficient at analysing data it cannot forge a rapport with individuals or empathise, which is essential in providing care.

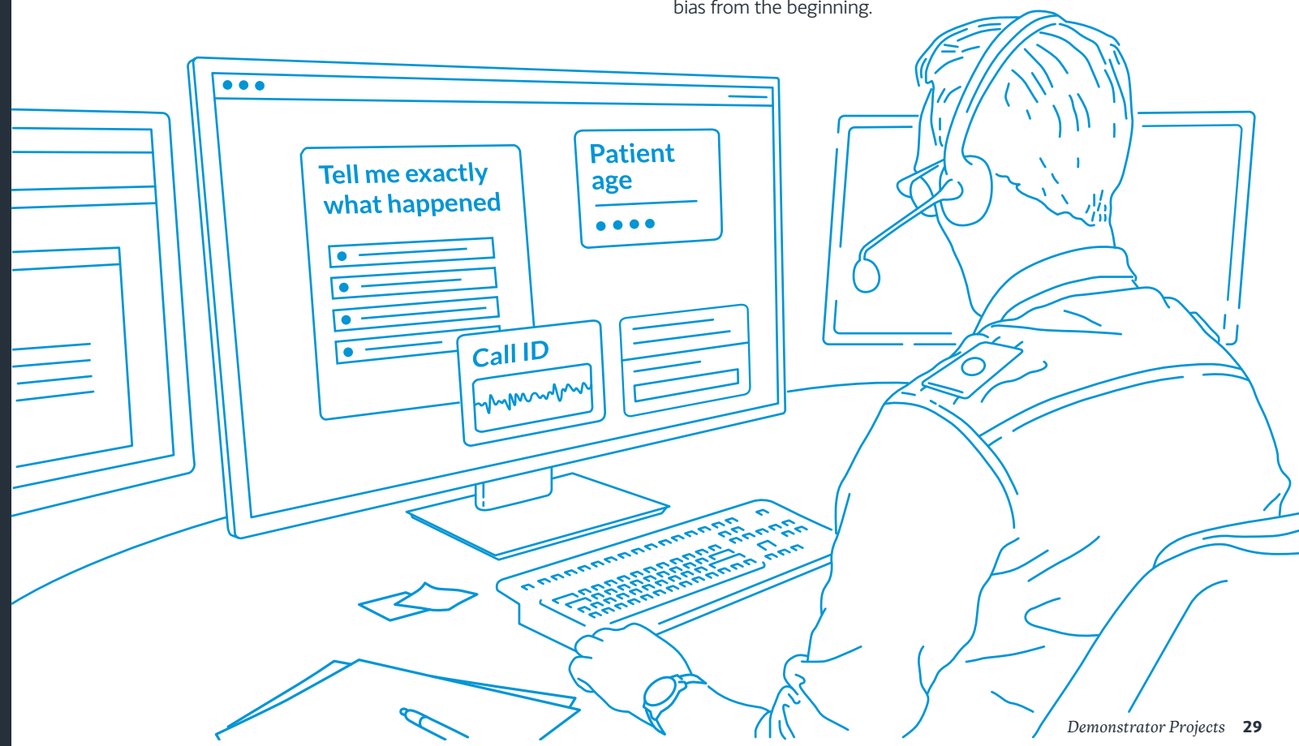
Creating and enforcing policy, regulatory and legislative mechanisms and ensuring the AI is ethical and legally justifiable is essential when integrating AI systems into healthcare settings. These mechanisms should specifically address risks such as automation bias and lack of interpretability and should be created, implemented, and monitored throughout all stages of the AI integration.

Looking specifically at interpretability, it is also essential to ensure the AI is fully explainable and that individuals working with or impacted by the systems are trained to understand how they work, especially with regards to their decision-making process (understanding how the AI came to a decision and identifying when the AI outputs may be wrong).

Whilst training is a key aspect of effective AI risk management, mitigating controls should be in place with regards to automation bias from the beginning. It is equally important to understand how AI systems can be useful to a healthcare service, to ensure that it is integrated into the correct departments and allocated to the correct tasks, as AI is not efficient in all areas. It can be invaluable in diagnosing and assisting, but cannot replace the experiences, intuition, and skills of HCPs.

Finally, it is essential to ensure that the AI systems are monitored (risk monitored) and analysed continuously when in use. For example, monitoring why and how frequently a decision or output made by the AI is rejected or accepted, in order to improve training and staff understanding of the systems, as well as monitor its effectiveness.

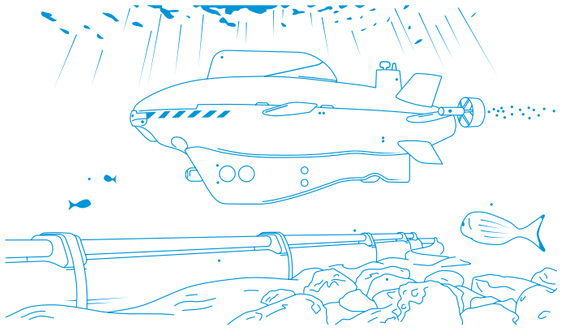
Dr Nigel Rees
Head of Research and Innovation
Welsh Ambulance Services NHS Trust



Assuring Long-term Autonomy through Detection and Diagnosis of Irregularities in Normal operation (ALADDIN)

Increasing the safety of marine autonomous systems (MAS) by developing a monitoring and classification tool that correctly detects and diagnoses unexpected vehicle behaviour.

What is the role of interoception in the assurance of autonomous systems, especially those that have long, remote missions?



An automated anomaly detection and fault diagnostic capability onboard an autonomous system reduces the possibility of it becoming a hazard to other systems and the environment. For instance, an underwater glider with power system failure or diving actuator failures can become an obstacle to other sea users, leading to increased collision risks with other systems (e.g. surface vehicles). A robust interoceptive capability onboard an autonomous system limits the possibility of itself becoming a hazard to a minimum by providing an immediate warning signal to the vehicle's safety and control modules. A remotely operated diagnostic system on the contrary, albeit more energy-efficient, is reliant on processing delayed decimated data transferred over satellite communication before being able to respond with appropriate preventive actions. The role of interoception is particularly significant for systems deployed over long-term remote missions where communication to the base station may be limited and where human intervention is challenging.

Additionally, an onboard system reduces the requirement of extended involvement of human experts. Diagnostics of problems can be limited by the experience of individual operators and the remedies are subject to human error. Therefore, reduced dependence on human expertise provides a higher level of safety and assurance of autonomous systems, provided that the interoception capability is robust in detecting critical failures. In the case of underwater gliders, which are typically deployed for remote missions over a long period of time, the current monitoring guidelines require the pilots' attention and to be present around the clock. This limits the scale of observational fleets that can be deployed simultaneously. Additionally, less reliance on human presence can significantly reduce the operational costs and enable larger scale simultaneous deployments of multiple autonomous systems.

Overall, generalised robust interoception systems onboard the autonomous systems can provide increased confidence in new autonomous technologies, increase in outputs, reduction in capital losses and operational costs, and even greater adoption of the new technology. In addition, such systems can also inform and improve the design and management of robotic and autonomous systems. For systems such as underwater gliders that are typically deployed for long remote missions, an automated and generalised onboard diagnosis protocol for adverse system behaviour through a real-time intelligent anomaly detection and fault diagnostics system can achieve a higher level of assurance; such that informed control actions can be made by the system's control and decision-making modules. Moreover, the records of the interoception system can be applied as a basis to update the system itself in the next design iteration, leading to even higher assurance of the autonomous systems.

Dr Yuanchang Liu
Lecturer in Autonomy
University College London

Boundaries Of AUTonomy (BOAUT)

Demonstrating and validating how onshore operators can detect and diagnose hazardous deviations by autonomous marine vessels in time for mitigating action.

What are the particular challenges in generating and validating operational scenarios in the maritime domain? Are all stakeholders (developers, operators and regulators) aware of their role in the generation of such scenarios?

The "fuzziness" of the domain is the largest challenge. It is not that there aren't rules, but traffic is mixed and practitioners acknowledge that rules are broken both quite frequently and for good reason. Looking at autonomy in particular, the challenge is that the roles are not yet set. As an example, in Sweden vessel traffic services (VTC) operators only provide advice that is not mandatory to follow. Whether this will change with the introduction of autonomous shipping is a large source of uncertainty. It is also quite urgent, as it seems autonomous shipping is quickly catching up with other domains, perhaps even leaving them behind soon.

I think all maritime stakeholders want to be involved in generating and validating operational scenarios, both for the sake of establishing assurance and for their own understanding. There is also a bit of a perspective shift taking place, in which VTC operators are thinking about new technology and how it fits into different situations. Either because they have seen it in other domains, or because they see challenges arising due to autonomy. However, the introduction of autonomous shipping seems to be right around the corner now, and some stakeholders seem to take the position that phased deployment could go hand-in-hand with detailed pre-analysis rather than waiting to follow it.

Dr Fredrik Asplund
Assistant Professor in Cyber-Physical Systems
KTH Royal Institute of Technology

CSI:Cobot – towards regulatory change

Shaping regulatory change using novel approaches to cobot safety.

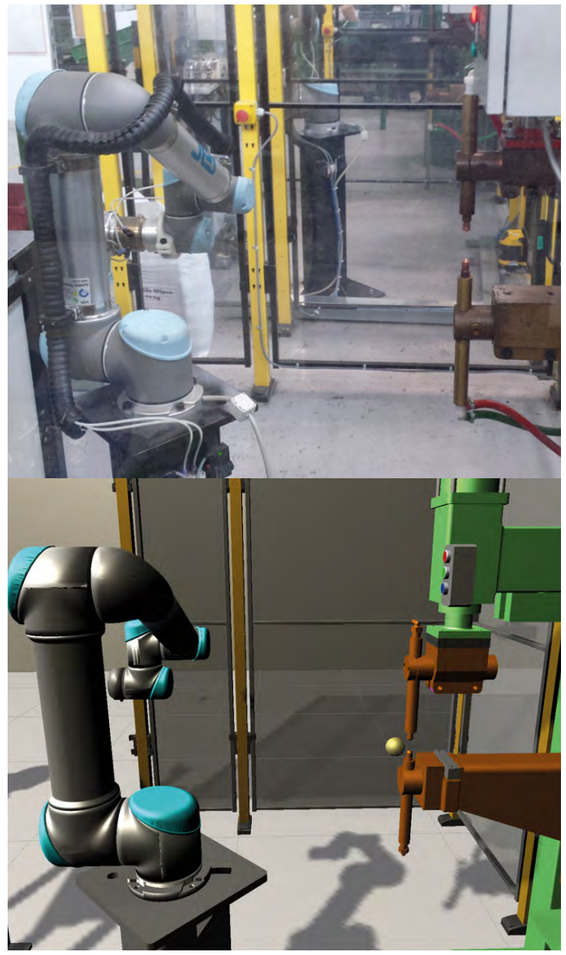
How can we help regulatory organisations keep pace with technology developments?

Regulatory organisations are often complex, including people with a range of technical competences, each with different roles and responsibilities. You will find scientists, engineers, inspectors, and others with specialist technical skills and knowledge.

To keep pace with advancing technologies and an ever-evolving landscape, regulatory bodies need to be agile and continually grow the skills and expertise they utilise. There are numerous ways to approach this challenge, but collaboration lies at the heart of the solution.

Collaborative research on specific topics can form the basis of partnerships with academia that benefit both organisations through the exploration of answers to new challenges. Pilot projects with industrial partnerships and organisations in the Catapult Network can help regulators to gain experience of novel real-world applications. Partnerships with other regulators, both at home and abroad, minimise the repetition of work that has already been completed.

These partnerships feed important information to the regulator, which builds internal competence and can be used to help develop policy and guidance and feed into the development of international standards.



CSI:Cobot phase 1 industrial case study robot cell and a representation of this in the project's digital twinning environment

Partnerships can also help regulatory bodies to achieve deep technical knowledge that would normally be achieved by in-depth training and time spent working on real system design. This approach wouldn't be suitable for day-to-day inspection and assessment activities but seeking advice from external organisations could help with reviewing new guidance or help with occasional accident investigation.

Through a collaborative approach we, in industry and academia, can support regulatory and standards bodies to build the skills, competencies and experience needed to solve the challenges of regulating autonomous technologies.

Dr James Law
Director of Innovation and Knowledge Exchange
Sheffield Robotics

Safe Airframe Inspection using Multiple UAVs (SAFEMUV)

Improving the safety of autonomous unmanned aerial vehicle teams through the creation of a systematic robustness assessment process.

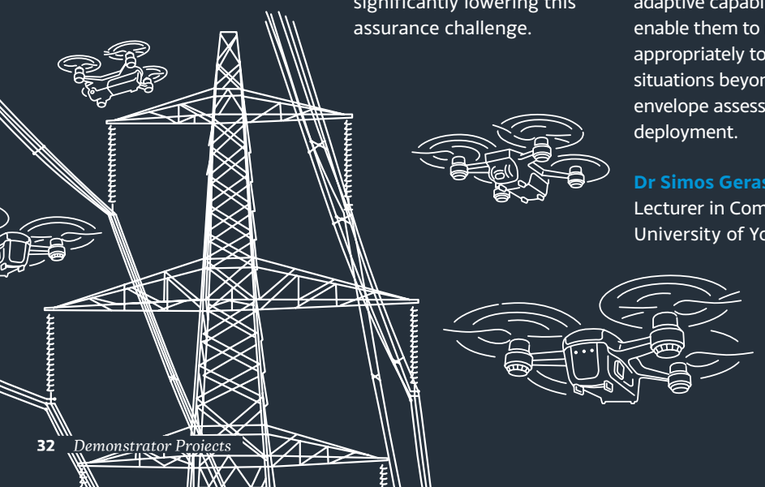
What are the particular assurance challenges of using drones for the inspection of high-value physical assets in a potentially harsh environment?

Employing drones to support the inspection of valuable physical assets is a highly sought-after innovation. It has the potential to improve the effectiveness of executing safety-critical missions, reduce the inherent risk of intervention from human operators, and make considerable contributions to environmental sustainability by reducing energy consumption and minimising pollution.

Notwithstanding the significant benefits and long-term societal impact of drones in inspection missions, their wider adoption has decelerated due to important and still pending assurance challenges. From a certification perspective, the lack of a complete, validated, and robust regulatory framework for UAVs is a significant barrier to ensuring and demonstrating their trustworthiness. The recently published pan-European UAV rules by the European Aviation Safety Agency is a promising step towards a harmonised regulatory framework for assured UAV operations. Nevertheless, the framework is abstract and still evolving. Furthermore, there is limited guidance on how regulatory bodies and operators should use it, especially for the most challenging mission types like inspection beyond visual line of sight using drones with a high degree of autonomous behaviour and reduced operator intervention. This challenge is exacerbated further because the publicly available documentation on how organisations and companies have applied this framework to assure their drones is quite sparse. Since intellectual property rights are involved, this is not surprising. However, adopting a more open and targeted dissemination strategy could help newcomers learn from best practices, thus significantly lowering this assurance challenge.

The uncertainty and openness of the environments in which the drones are typically deployed constitute another fundamental assurance-related challenge. Unavoidably, drones do not operate in 'Faraday cage'-like settings but within shared environments where humans, animals, or other static or mobile robots reside. These environments may also involve partially or fully unknown segments where the behaviour of actors or the environment itself cannot be controlled or is not entirely understood. Since anticipating all potential situations that the drones may encounter before deployment is impossible, we can only provide partial assurances at design time. Thus, the assurance case needs to be a living and continually evolving artefact, updated when new evidence becomes available at runtime. Consequently, engineering drones with these capabilities that can also operate with increased trustworthiness levels in these environments requires assurance-informed context-awareness and adaptive capabilities that enable them to respond appropriately to emerging situations beyond the safety envelope assessed before deployment.

Dr Simos Gerasimou
Lecturer in Computer Science
University of York



The Thorvald robot used in the project

Medium-Sized AGV for soft-fruit Production (MeSAPro)

Assuring the safety of autonomous robotic systems that will support human fruit pickers and reduce workplace accidents.

Agricultural robots often need to work in close proximity to humans. How do we deal safely with the challenges this brings?

We are working on the technical aspect of ensuring that robots can safely work in close proximity to humans. This starts by defining those situations that are safe and those that are not. We can then program the robot to identify an unsafe situation and to take the necessary action to return to safety.

For example, in soft-fruit production robots are used to treat plants with high-intensity UV light to kill powdery mildew. This UV light can hurt people – in essence giving them a bad sunburn – if they move closer than 7m from the robot while the UV lights are on.

The robot is therefore programmed to detect if someone is closer than 7m from it and to immediately shut the UV lights off if this is the case. The lights are not reactivated until an operator has checked that the area around the robot is clear.

Robots are also used to transport picked fruit around the farm. To do this, they have to approach a fruit picker closely enough that the picker can place trays of fruit on the robot. Previous research has established that people react differently to robots further than 3.6m away (the "public zone"), between 3.6 and 1.2m away (the "social zone") and less than 1.2m away (the "intimate zone").

The robots are programmed to recognise when they are approaching the edge of these zones and to change behaviour to respect them. Accordingly, a robot will slow down and stop as it approaches the boundary between the public and social zones, and will only proceed when a person signals to them that it is ok to do so. The robot will then move slowly towards a picker across the social zone until it reaches the boundary of the intimate zone, where it will stop. It will not move any further towards the picker, but, when the picker signals to it, will then move away.

There is another, equally important, regulatory aspect to the safety of agricultural robots. This involves the updating of rules around the use of farm machinery to deal with the autonomous nature of robot operations. To support this process we are engaging with the British Standards Institution.

Professor Simon Parsons
Head of School of
Computer Science
University of Lincoln

Autonomous Capabilities and Trusted Intelligent Operations iN Space (ACTIONS)

Improving the utilisation of satellites for data capture through the safe introduction of autonomy.

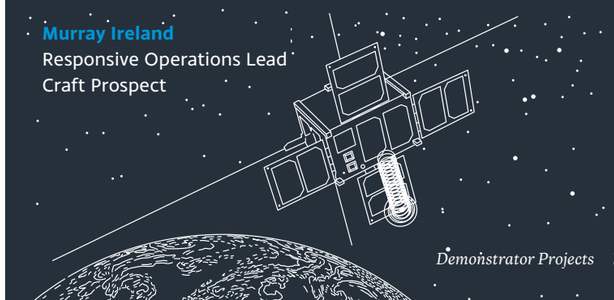
What are the particular assurance challenges of working on space systems that have such tight constraints? Does this provide lessons for other domains?

Space provides a novel and challenging environment for the deployment of a robotic system. The ability to test a space system in a fully representative environment before launch is impossible, and so tools such as hardware-in-the-loop simulation must be heavily relied upon. This is only compounded in the recent developments in autonomous space systems. In many cases, such autonomy is driven by visual data acquired from on-board optical instruments, passed through neural networks to extract information about features on the Earth and in the local orbital environment. Training data for such activities is very limited, particularly when considering the specific features of interest and properties of the capturing instrument. This, combined with the limited processing power of on-board computers, restricts the performance – both accuracy and speed – of any neural networks deployed on-board. This has a subsequent impact on the safety of the autonomous satellite, to itself, its data (a key concern of end users), and even to life on the ground in some applications.

As satellite autonomy is still in its infancy, the tools and resources required to adequately assure the safety of an autonomous satellite are relatively immature. We have spent much time defining and characterising a representative mission for investigation during our AAIP project, ACTIONS. We are now simulating this mission with representative flight hardware in the loop, so that we can truly understand the impact of low-power neural network performance on the mission behaviour and results. This simulation requires a close coupling of orbital mechanics, spacecraft dynamics and visual inputs to the simulated instrument. It is only by doing this work that we can test the neural network in a realistic context

Such work has lessons for other domains, primarily those where testing in the intended operating environment is infeasible or impossible, such as deep-sea missions or robots used in the nuclear industry. In these cases, simulations of sufficient fidelity can be used to model the operating environment and provide the relevant stimuli (image data and environmental disturbances) to fully test the autonomous system's response and be confident in its safety.

Murray Ireland
Responsive Operations Lead
Craft Prospect



The Future

Our guidance is already influencing the way developers and others approach the safety assurance of autonomous systems. In the year ahead, we will continue to publish new methodologies to further support those working in this area.

With Assurance of Machine Learning for use in Autonomous Systems (AMLAS) currently used by colleagues across the world, we are looking forward to launching an interactive resource, case studies and supporting tools to complement this leading guidance and make it more accessible and easier to apply. We will also publish our Safety Assurance of autonomous systems in Complex Environments (SACE) methodology, once it has completed its final round of peer reviews.

AAIP was the cornerstone of the University of York's successful bid for funding to establish the Institute for Safe Autonomy (ISA). The Institute formally opens its doors in 2022 and the purpose-built centre will become the new home for our established and continuing research. The new facility will bring together researchers from across the University to work on aspects of autonomy and supporting technologies, with our research forming the core of the Institute's assurance pillar.

The laboratories available to us in the Institute will offer new avenues for collaborative research and we are already in conversation with partners about how we can advance our work on assurance and validate concepts with the space and technology available.

As well as working in the new laboratories, we plan to use the space for public engagement, building on our existing work to understand people's perception of autonomous systems and increase awareness of the technology. We are already working with the Science Museum Group to develop an exhibition on autonomous technologies for summer 2022. The ISA building offers other chances for people to interact with the technology to improve their understanding of it and we look forward to maximising these opportunities.

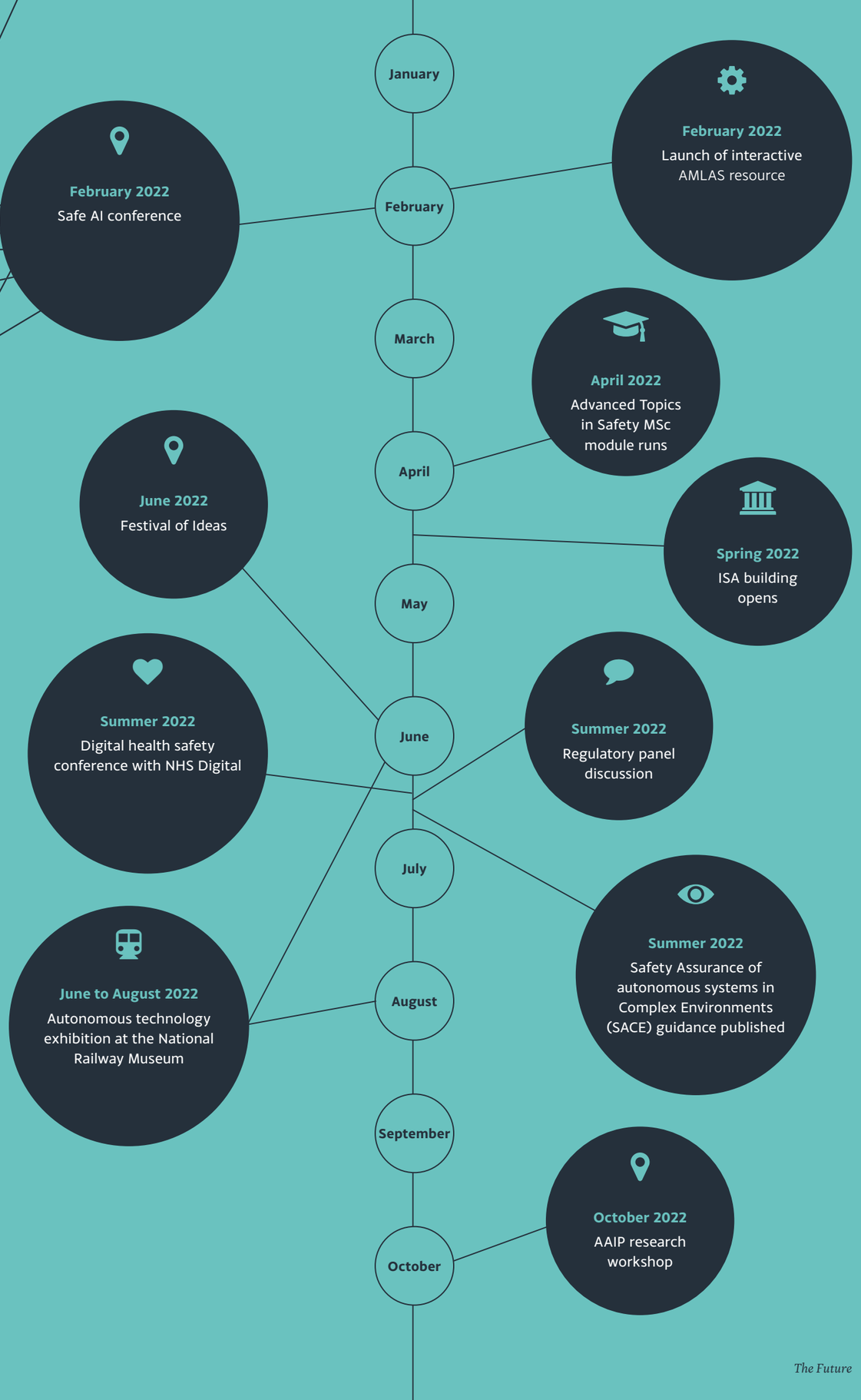
As we enter the penultimate year of the programme we can see the influence our work is having. Through collaborations with industry, the involvement of regulatory and standards organisations in our research, and growing engagement with the public we will ensure that we continue to impact the development, regulation and wider understanding of autonomous technologies.

Work with us

The challenge of guiding the safety assurance and regulation of autonomous systems is something that must be done in collaboration. It requires further funding, joint research, evolving regulations, and more. Contact us if you would like to collaborate to ensure that autonomous systems are designed, developed and introduced safely to benefit us all.

+44 (0) 1904 325345
assuring-autonomy@york.ac.uk

@AAIP_York
linkedin.com/company/assuring-autonomy
assuringautonomy.medium.com



UNIVERSITY of York

Lloyd's Register Foundation