

A YEAR IN REVIEW 2020



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

Key

Funders
 Demonstrator projects
 Collaborative links
 Programme Fellows



Contents

Welcome	4
A year in numbers	6
International Community	8
Public Engagement	10
Education and Training	12
Body of Knowledge and Foundational Research	14
Advancing regulations	16
Plans into policy: the journey of healthcare regulations	18
Demonstrator Projects	22
The Future	34

Welcome



This year has been quite unexpected in many ways. The COVID-19 pandemic has affected us all. It has changed the way we work while bringing additional creativity and strength to our collaborations.

nen I wrote this welcome letter for our 2019 review I could never have imagined what lay ahead in 2020. But we have accomplished much of what we planned and achieved additional successes along the way.

Again our team in York has grown. We have strengthened our work in the ethics and governance of autonomous systems, expanded our technical expertise, and enhanced our industrial experience through the recruitment of new research associates and engineers.

We have increased our funding of real-world research by commissioning an additional £1.2M of demonstrator projects. These projects span a variety of sectors and are all working with regulatory bodies to move their work from safety processes towards the introduction of standards and regulations.

We continue to further our standing as the leading centre of excellence for the safety assurance of robotics and autonomous systems. The new facility that will house the Institute for Safe Autonomy is taking shape on the University's Campus East and will bring together our work with

complementary University expertise including robotics, advanced communications and quantum technologies, and enable greater industry and regulator involvement.

An award of more than £3M from the UKRI Trustworthy Autonomous Systems (TAS) programme, has enabled us to establish and lead the TAS Node in Resilience. This node brings together multidisciplinary capabilities from five UK universities to improve the ability of autonomous systems to reason about the impact of their decisions and actions on technical and social requirements and rules.

This year has clearly demonstrated just how much we will all benefit from the introduction and adoption of autonomous systems, and the work we're doing to ensure this is done safely is more important now than ever before.

We are proud of what we have achieved this year and look forward to continuing global, multidisciplinary collaborations in 2021 to ensure safe autonomous systems are a reality for us all.

Professor John McDermid OBE FREng **Programme Director**



2020 has been a year unlike any we have known in living memory where the COVID-19 pandemic has brought the understanding of risk into public focus. The public understanding of risk has been an important area of Foundation activity for several years. Our World Risk Poll has started to track global attitudes to risk and what people across the globe really worry about and what they have experienced. Through the poll we have started to build a picture of people's views on topics including climate change, food safety, the digital world, and their health, safety and wellbeing at work.

The introduction of autonomous The Programme connects the

technologies has the potential to reduce the risks that people are concerned about, however, their introduction is not without risk itself. Our partnership with the University of York has created a global resource for ensuring autonomous technologies are introduced safely. brightest minds in industry and academia to together solve the challenges of assuring the safety of autonomous systems: learning from multiple sectors to create solutions which will stand the test of time and be usable across domains and across geographic boundaries.

Professor Richard Clegg FREng Chief Executive Lloyd's Register Foundation



The University of York exists for public good. The work of the Assuring Autonomy International Programme supports this vision through the creation of strong partnerships: partnerships built on a shared commitment to the safe introduction of autonomous technologies.

Established from a collaboration between the University and Lloyd's Register Foundation, the Programme connects researchers, developers and regulators to start to agree the new principles and rules for safe AI, robotics, and autonomous systems.

By working in partnership across the globe, across disciplines and sectors we can help industry to prove that their systems are safe, help regulators to set consistent safety standards, and empower people to be confident users of the technologies that are increasingly evident in our lives.

Professor Kiran Trehan

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





Minimising and managing system failures workshop, London, February 2020

International Community

This year we have seen changes in our way of working, but we have capitalised on online networking and collaboration tools to enable us to form new global partnerships and strengthen existing ventures.

> arlier this year we launched our online Body of Knowledge and have been pleased to find that its guidance has been accessed thousands of times by people across at least three continents. We have also continued to work across the globe through our research, Fellowship scheme and other networks.

Europe

We now have 11 demonstrator projects in place across European countries,

totalling more than £2 million of funding. These projects cover a range of domains including agriculture, automotive, healthcare and manufacturing.

We are delighted that Dr Radu Calinescu is the Principal Investigator on an EPSRC-funded Trustworthy Autonomous Systems Node in Resilience. The project will fund a multidisciplinary team across the five universities involved to develop a comprehensive toolbox of principles, methods, and systematic approaches for the engineering of resilient autonomous systems.

We have made good headway in

the automotive domain this year. In particular we launched "Assuring the safety of highly automated driving: state-of-the-art and research perspectives", written by AAIP Fellow Professor Simon Burton and Dr Richard Hawkins. We continue our partnership with Five, working with them on perception research, and are pleased to be part of Zenzic's UK Connected and Autonomous Mobility (CAM) work as a CAM Creator.

We continue to build on our strengths in the healthcare domain. Dr Chiara Picardi and Dr Ana MacIntosh were both part of the Assuring Safety for Assistive Robots in Health and Social Care workshop at the European Robotics Forum (ERF) in Malaga, Spain, led by our partners Bristol Robotics Laboratory. In spring we had a paper published in the World Health Organization Bulletin. "AI in healthcare: accountability and safety" considers the possibility of patient harm caused by the decisions of an AI-based clinical tool, focusing on two aspects: moral accountability and safety assurance. We are also collaborating with clinicians, statisticians, patient groups, and patient safety and health innovation experts on a project to evaluate the safety of a pioneering AI healthcare Command Centre in a Bradford hospital. As well as funding a new

demonstrator project, SAFEMUV, which is undertaking research into the assurance of autonomous unmanned aerial vehicles that carry out safetycritical inspection tasks, we also continue to collaborate with the European Robotics for Inspection and Maintenance (RIMA) Network on the safety and certification of inspection and maintenance robotics. Most recently we launched a new report "Review of legal frameworks, standards and best practices in verification and assurance for infrastructure inspection robotics", which provides a single point of reference on the safety, regulatory and



Dr Ana MacIntosh at ERF, Malaga, March 2020

liability issues for operating inspection and maintenance robots in the EU.

Asia

The TIGARS demonstrator project concluded earlier this year. The team made good progress in the assurance of autonomous vehicles, as well as headway in the international standards arena through work with regulators in both Japan and the UK.

We were pleased to welcome new Programme Fellow Tarek Nakkach to York earlier this year to introduce him to the AAIP team and other collaborators at York. Tarek is Regional Legal Counsel for UKIMESA Hewlett Packard Enterprise. His multidisciplinary work is looking at issues around the liability and ethics of autonomous systems and he has written a number of blog posts for AAIP on the topic.

Australasia

The ATM demonstrator project continues at the Australian National University, with preliminary evaluation of the testing mechanism and software framework on a high-fidelity autonomous car simulator underway.

with the Trusted Autonomous Systems Defence Cooperative Research Centre to consider how we might work together on areas such as training, research, and Body of Knowledge guidance.



AAIP research workshop, online, October 2020

We are continuing our conversations

North America

We have been involved in a number of conferences and events linked to partnerships in North America. In February Dr Colin Paterson presented our AMLAS (Assurance of Machine Learning for Autonomous Systems) process at the Safe AI workshop in New York. Since then we have been refining and validating the process to develop a practical handbook for safety engineers and developers.

Dr Richard Hawkins presented at the 10th annual ISO26262 conference, which was held online instead of in the USA. His presentation considered the challenges to the safety assurance of autonomous driving.

We were delighted to host an experts' workshop to prepare for the first International Congress for the Governance of AI (ICGAI). Working with Congress leader Wendell Wallach from Yale University, we brought together experts from across the globe to discuss minimising and managing system failures.

South America

We have continued to work with Fellow Genaína Rodrigues who has investigated research challenges associated with the modelling, verification, planning and simulation of multi-robot systems. In collaboration with her graduate students and Radu Calinescu's York AAIP team, she is developing a goal-oriented modelling approach for robot mission specification. Genaína and her research group are also developing a lightweight simulation environment capable of handling larger multi-robot systems than those supported by traditional simulation environments, such as Gazebo.

More information about working with us: bit.ly/aaipworkwithus

Public Engagement

The use of artificial intelligence (AI) and machine learning is more evident every day: banking systems that manage your money, predicting when and how much to save; TV that recommends what to watch next based on previous viewing. We might be happy to use these advisory systems, appreciating the benefits they bring. What if this AI or autonomy was used in a safety-critical situation such as in a car or a hospital? Would we want to know more about the risks?

nis year our public engagement work asked that question. We held five focus groups to talk with people about their opinions on technology and autonomy generally and in relation to two specific scenarios: autonomous cars and healthcare.

Attitudes towards technology

The groups we spoke with mostly enjoyed the benefits of technology, such as saving time or enabling them to connect with family and friends across the world. They could also see benefits for wider society, for instance more environmentally-friendly transport options or improved medical treatments.





You know, shopping. It's saving us having to go out, so we've got more time for ourselves. Even booking an appointment at the doctors now I do it online."

They had concerns about the impact of an increasing use of technology in our lives. These were mostly around an over-reliance or an intrusion of technology, and a worry that the technology could go wrong. Some were also concerned that technology moved too fast and was more for company profit than to help society.

Are we still the master of the technology? Or do you develop a robot or a program that does the thinking for us?"

The things that I find too fast are things that I feel are driven by capitalists where they're just trying to get you to buy the new model and old ones become obsolete."

Autonomous vehicles

"It's acceptably safe" is how one member of a focus group described driving a car at the moment. "It's clearly not safe, but it's a means of getting around so I'm more than happy to take the risk," they continued.

When considering autonomous vehicles, the groups divided into three main categories - supporters, potentials and rejectors.

Supporters suggested that the technology can outperform humans, would maximise leisure time and would be less stressful than driving.

ABS does a pump 500 times a minute. We can't pump brakes as fast as that. So there is an example. Technology is working and a car would react and see a person *auicker than we can.*"

Potentials want more information. They question the technology's capabilities and want to know more about the risks and who would be liable in the event of an accident.

It's like when you take your driving theory test and particularly when you see the hazards, they're not actually something happening yet but as a human you can say, right, I think that person might be thinking about crossing the road, where a computer couldn't necessarily judge."

Rejectors were very sceptical, had a mistrust of the technology and were not always keen to listen to more detail about it.

I was going to say, they'll have to be out for about ten years before I'd trust it. After that maybe."

The groups saw two key barriers to the introduction of autonomous cars: infrastructure and the level of autonomy offered.

autonomous vehicles alongside current cars, with groups worried that until the infrastructure is in place to fully support autonomous cars, the mix of old and new technology on the road would be riskier.

I consider the most dangerous time will be the interface between a mix of autonomous vehicles and when other drivers are still on the road."

Cars which still require a safety driver to be alert at all times were thought to be "pointless". Some of the people were onboard with autonomous vehicles, but they wanted to be able to hand over full control to the system.

Was the driver expected to be paying attention to what was going on outside the car while it was in full computer mode? If it was, what's the point of full computer mode?"

Healthcare

To demonstrate the use of AI in healthcare, the groups were given

There was an unease about the use of

If I got an autonomous car I'd want to be chilling in the back." a scenario of a hospital ward using an autonomous infusion pump to provide medication.

The discussions highlighted an inherent trust in the healthcare sector, with wide acceptance of the use of the autonomous infusion pump, and a clear understanding of the benefits.

They're professionals, they know what they're doing and, yeah, [I would] completely accept it."

So I think we just have to assume that they've got systems and controls in place within their organisation to make sure that the machine works."

I think if it would save health professionals time, like nurses, then it does sound like it would be a very good thing, if it was reliable enough."

The groups did still have reservations about the use of the technology, but generally their verdict was much more positive than in the automotive scenario.

I'd still be a little concerned... If it was one computer on one person not a problem, because that's what it is basically now. But if it was ten *people* [monitored by one computer] there is still a chance that something could go wrong."

Next steps

The focus groups enabled us to distil some key themes that we will take forward in 2021 to develop a much larger study of the public's perception of the risks of autonomous technologies.



Education and Training

This year we focused on the key audience of developers and safety engineers. We have created a range of opportunities for them to develop their skills, knowledge and behaviours.

s some of our demonstrator projects come to an end, we have seen an increase in high-quality case studies and practical guidance informed by their real-world experiences. This has enabled us to focus on developing three key education and training opportunities for those developing autonomous technologies.

Academic education

In September 2020 we launched the development of Advanced Topics in

Safety - our new 10-credit module for the University of York's MSc in Safety-Critical Systems Engineering.

This module is our first formal academic education opportunity in the safety of autonomous systems. It represents an opportunity for students to broaden their system engineering knowledge with an insight into the challenges to safety processes and product safety and initial approaches to addressing the challenges of autonomous systems.

Over the coming year we will build on our academic education offer with further embedding of material into the



Dr Mark Nicholson presenting to clinicians and other staff at NHS training, January 2020



modules of the MSc in Safety-Critical Systems Engineering and advice to academic education providers on how to embed safety into their robotic and autonomous system courses.

Industrial education

In partnership with NHS Digital we launched a new training programme for those working in health IT and

Clinical practice and the efficiency of healthcare provision will benefit from the introduction of AI and autonomy, but patient safety must still be our priority. Understanding these technologies and how they could change the clinical setting is important for everyone within our organisation. The training workshop provided by the AAIP was delivered perfectly for our audience of clinicians and other NHS staff and the feedback has been very positive."

Sean White, Senior Safety Engineer, NHS Digital

clinical risk management to give them an understanding of the use of Artificial Intelligence (AI) and Machine Learning (ML) in healthcare.

The course builds on previous workshops run for clinicians in partnership with NHS Digital, and we will deliver more in 2021. It covers: introduction to AI and ML

- the challenges of AI and ML in health and social care
- assurance of ML
- safety requirements and data management for ML
- training and verification of ML

Over the coming year, we will build on our professional development education offer by working with industrial organisations to develop and embed robotics and autonomous system safety engineering education into their competency and professional development schemes. We will offer a series of webinars to the AAIP community.

Developing young safety professionals

We are delighted to be working with the Safety-Critical Systems Club to launch the Safety Futures Initiative (SFI), for young and early-career safety professionals. The initiative brings together early-career professionals to:

- gain experience
- share knowledge
- specialised interest area of safety engineering and assurance

safety cases for ML in a clinical setting

build a community around the

As a young professional it can be difficult to find opportunities to learn about system safety. This initiative will build a community of likeminded individuals to share ideas, experiences and develop together in the field of system safety. As we move more and more to a world of autonomous systems we need to ensure that those who will be *developing and working* with these complex and often safety-critical systems have the knowledge, skills and behaviours that are needed."

Nikita Johnson, SFI Lead

We could develop a bespoke training course for your organisation: bit.ly/aaipeducation

Body of Knowledge and Foundational Research

Our research focus this year has brought together the practical guidance developed for the Body of Knowledge with our foundational research on the key challenges that remain at the heart of the safety assurance of autonomous systems.



ver the year, we have added to the practical guidance in the Body of Knowledge with expert advice from our demonstrator project teams and other AAIP collaborators. This has been accessed from across the globe by the key audience of safety engineers and leaders.

We have also been advancing our research on some of the core technical issues that remain for the safety assurance of autonomous systems and have established a new structure to focus our work. This includes five key pillars of research:

- 1. assurance of machine learning in autonomous systems (AMLAS)
- 2. safety assurance of understanding in autonomous systems (SAUS)
- 3. safety assurance of decision making in autonomous systems (SADA)
- 4. safety assurance of autonomous systems in complex environments (SACE)
- 5. social acceptability of autonomous systems (SOCA)





We are developing a manual for each of these areas. These will be based on sound research and empirical evaluation.

The guidance provides practical notes, examples and links to the Body of Knowledge, creating a complete handbook for safety engineers, developers and regulators.

More than the sum of its parts

Each guidance document will stand alone as an essential component for assuring that aspect of an autonomous system. However, assurance of autonomous systems must consider all aspects in the broad context of development, operation and approval. Used together, the guidance documents will help to ensure a credible and compelling assurance case is created for an autonomous system.

Assurance of machine learning

Our published AMLAS process (Assurance of Machine Learning for use in Autonomous Systems) has been developed into our first guidance document. This has been peer-reviewed by experienced engineers from multiple industry domains and our Fellows (who had a first glimpse of the new process).

AMLAS comprises:

- a set of assurance activities that integrate with the development of ML components
- defined assurance artefacts relating to those activities
- safety case patterns to guide the development of a compelling safety case for ML components

The integration of activities, artefacts and safety case patterns ensures AMLAS provides a practical and coherent approach.

Using AMLAS

We are currently creating an online resource for all of the guidance documents. This will provide an interactive, accessible way to use them.

In the interim, we have published an interactive PDF of AMLAS which is available to download for free from our website.

Download the AMLAS: bit.ly/aaipamlas

Advancing regulations

Regulations and standards could be the make or break point for autonomous systems. The technology developers continue to make remarkable progress and the AAIP and others are introducing practical guidance for the assurance of autonomous systems. Yet without appropriate rules and principles in place to guide their introduction and use, we may end up with technology that cannot be safely deployed.

complex and interwoven. Each domain and country is different. Many sectors have numerous regulatory bodies and countries across the world enforce compliance of regulations in different ways, making progress slow and challenging.

he regulatory landscape is

In 2020 we started our work with regulators to move us towards the introduction of new principles for safe AI, robotics, and autonomous systems. The COVID-19 pandemic has slightly impacted our progress, but through collaborations with partners and demonstrator projects we have begun work in numerous domains. We know that we must collaborate to advance this work and 2020 has given us an opportunity to make connections where needed.

Healthcare

We have made good progress in the healthcare domain, through both demonstrator projects and the work of our researchers and collaborators.

In February, we worked with NHS Digital to bring together experts from the health and social care regulatory bodies to discuss a whole system approach to the safe introduction of AI in the NHS. The discussion was focused at the system-level; considering development, assurance and deployment of AI tools as closely interconnected elements. We examined and evaluated existing processes for software safety governance in the NHS and considered how to support the use of AI.

A demonstrator project, led by NHS Digital and other regulatory partners,





has been funded by the Programme. This project (Safety Assurance FRamework for Machine Learning in the Healthcare Domain) will produce concrete guidelines for the safety governance of autonomy in the NHS.

The SAM project (see page 24) has been involved in setting up the Digital Health and AI special interest group of The Chartered Institute of Ergonomics and Human Factors (CIEHF). The group aims to provide human factors guidance on AI in healthcare to regulators, developers and users of AI and digital technology in the health sector.

Maritime

We have been working with BMT and Maritime UK to identify ways in which we can impact the governance of this complex and interconnected sector as the introduction of autonomy becomes more widespread. In 2021 we will hold a workshop with key stakeholders from maritime and beyond to assess the challenges in governing the movement of goods from one UK port to another, when autonomy and AI are applied in various ways either on the vessels themselves or through the use of autonomous vehicles portside.

Automotive

Through numerous regulatory workshops in the UK and Japan the TIGARS project (see page 22) furthered work in the international standardisation area. In particular, Kanagawa University has used the project's standardisation activities to support revisions of numerous ISO/ IEC/IEEE standards.



We also published "Assuring the safety of highly automated driving: state-of-the-art and research perspectives", written by Professor Simon Burton, one of our Programme Fellows, and Dr Richard Hawkins. This new report presents a framework for assuring the safety of highly automated driving systems, which is the first step if we are to be able to regulate these vehicles.

Plans into policy: the journey of healthcare regulations

Healthcare has an array of systems and tools that make use of artificial intelligence (AI) or machine learning (ML). Some have been approved. For example, a Deep Learning Image Reconstruction system, developed by GE Medical Systems, has been approved by the FDA in the US. While healthcare standards and regulations are in place, challenges lie in the suitability of these to provide effective oversight of such technology. In 2020 we worked with UK health and social care regulatory bodies to make headway with the resources and exemplars needed by regulatory bodies to support the safe introduction of AI and ML in healthcare.

QUARTER ONE, 2020 Identifying the challenges

In January 2020 NHSX convened a senior-level roundtable discussion with organisations involved in the regulation of AI in healthcare in the UK. The group identified several actions needed to enable the UK to become a world leader in AI in health, including using a joined-up approach to the work needed.

To move this group's findings forward we worked with NHS Digital's Clinical Safety Team to assemble a second roundtable discussion.

This brought together NHS Digital, NHSX, the Medicines and Healthcare products Regulatory Agency (MHRA), NHS National Services Scotland, the Care Quality Commission, the British Standards Institution, NHS England, NHS Improvement and others to discuss a whole system approach to the safe introduction of AI in the NHS.

The primary conclusions from the discussion focussed on what we need to ensure we can develop an integrated safety governance framework to support the use of Al and ML in the NHS:

- clear, practical guidelines
- skills, training and resources
- clear connections between the different regulatory and assessment agencies
- approval mechanisms for updates
- clear evidence requirements
- post-market surveillance

The group committed to working together to produce concrete guidelines for the safety governance of AI in the NHS through the development of a demonstrator project to be funded by AAIP.



QUARTERS TWO AND THREE, 2020 Project development and approval

NHS Digital worked with the organisations from the roundtable discussion, along with manufacturers of healthcare technology, hospitals trusts and others, to lead the development of an AAIP demonstrator project proposal.

The group had identified the need for robust methods to assure the safety of machine learning systems in healthcare and wanted to develop a project that would help to establish a safety assurance framework to support healthcare manufacturers and deploying organisations to:

- assure their ML-based technology
- meet their regulatory obligations

They recognised the need for wide stakeholder involvement to ensure the project outputs worked in practice, were informed by needs on the ground, and would lead to the necessary changes to policy and practice. They also wanted to include consultation with patient groups in order to establish what assurance and confidence patients would expect to see before such technologies are used to deliver care.

The project developed and approved by AAIP governing bodies – Safety assurance framework for machine learning in the healthcare domain (SAFR) – is underpinned by our AMLAS work, and will develop a machine learning safety assurance process model supported by healthcare domain tailored artefacts and exemplars.

QUARTER FOUR, 2020 **Project kick-off and** introductory work

The SAFR project started in the last ouarter of 2020. The team has started work with the literature review to determine the scope and adequacy of existing resources to support successful regulation of ML technologies in the healthcare domain.

Safety assurance framework for machine learning in the healthcare domain (SAFR)

Creating resources to help manufacturers and others to meet the regulatory requirements for their ML healthcare tools

This project will use a literature reviews, expert workshops and case studies to build an evidence base to support the development of a safety assurance framework for the use of ML in healthcare. It will answer three key research questions:

- What published literature exists that aligns with the requirements of AMLAS?
- Does a specific instance of the AMLAS need to be established for healthcare?
- Can AMLAS be applied in practical ML-enabled healthcare systems and support compliance with the associated regulation?

The project aims to publish outputs as guidance for regulatory bodies, manufacturers, and user communities, to support their implementation and regulation of ML systems, and to influence the development of future regulations.

- **Project partners**
- NHS Digital
- British Standards Institution
- Human Factors Everywhere



2021

An invitation to work with us and replicate the process in your domain

We are already in the process of following this successful process in other domains. We welcome discussions from regulatory bodies and industry partners who wish to work with us to progress regulations and standards in other domains.

Please contact Sarah Heathwood, Partnership and Communications Manager: assuring-autonomy@york.ac.uk

Demonstrator Projects

People need to have confidence that autonomous systems will do what they expect. That might be the engineer designing a new feature, the safety expert validating an edge case, the operator interacting with the system, the public whose lives could be impacted, or the insurer that needs to foot the bill if something goes wrong. ur demonstrator projects all progress the assurability of robotics and autonomous systems (RAS) in different domains. They contribute pragmatic, evidenced techniques that can be used by developers, regulators and others to ensure the systems they develop are safe for users and others who come into contact with them.

This year we invested a further £1.2M in four new projects, across a range of new domains including maritime and agriculture.

More information about demonstrator projects: **bit.ly/aaipdemonstratorprojects**



Safe-SCAD driving simulator (photo by Dan Addison/UVA University Communications)

Safety of shared control in autonomous driving (Safe-SCAD)

How do humans and machines safely share control of an autonomous car? The goal of the Safe-SCAD project is to deliver methods for assuring the safety of shared control in semiautonomous driving.

The team used a highfidelity driving simulator to collect data about a human's response time and physiological signals in different driving scenarios. From this they developed DeepTake, a novel deep neural network-based method that predicts the driver's intention (at 96% accuracy), time (at 93% accuracy), and quality of takeover (at 83% accuracy).

The Safe-SCAD project targets important barriers that prevent the use of shared control in automated vehicles, and thus has the potential to make a significant contribution to their safe adoption."

Lu Feng, Assistant Professor of Computer Science, University of Virginia

The team investigated the applicability of formal methods to neural networks. They applied the Marabou verification tool for local robustness checks. They also employed clustering to find regions in the input space of the neural network where the network is robust. Furthermore the team has investigated three stateof-the-art tools, namely LIME, SHAP and Integrated Gradients, for explanation in neural networks, and has applied them to the "Takeover-time" neural network. The result of the analysis is a set of high importance input features that can be used to understand and certify the network; they can also be used to compress the network.

The team has defined the inputs, actions and requirements for a handover safety controller that will exploit the predictions provided by the neural network model and its verification results, and provide a driver availability recognition and improvement system.

The defined requirements will reduce the frequency of minimum risk manoeuvres (e.g. unplanned stops due to safety concerns) by issuing auditory/visual/ haptic warnings to improve driver availability and/or by suggesting speed adjustments to the car controller.

For the next step, the team plans to implement the Safe-SCAD controller on the driving simulator and conduct more human subject experiments to evaluate the performance of Safe-SCAD system. The results of the formal verification will be used in the design of the controller. The project team will also reach out to regulators and industry partners to discuss the insights learned in this project.

Project partners

- University of Virginia
- Carnegie Mellon University
- University of York



Collecting off-road data

Sense-Assess-Explain (SAX)

How can we develop autonomous vehicles that can explain the decisions they take?

This project is designing, developing, and demonstrating autonomous vehicles that can:

- sense and fully understand their environment
- assess their own capabilities
- provide causal explanations for their decisions

The team at the Oxford Robotics Institute (ORI) has been researching and improving the efficacy of traditional (e.g. cameras and lasers) and non-traditional (e.g. radars and microphones) sensors for understanding the world around the vehicle, including dynamic objects and drivable surface as well as measuring the vehicle's movement and location.

In the lab, the team is turning its attention to using this robust sensing to enable the vehicle to share and explain its observations of the environment in human-understandable terms via models of run-time capability.

The next steps are to expand off-road trials to new sites and diversify a planned dataset with urban, suburban, and motorway driving conditions.

Project partner

• University of Oxford

In this project, we have tried to push autonomous driving to its limits, in scenarios where untrained people could struggle. Trying to enable autonomy in these scenarios is forcing us to develop innovative techniques, showing that the field is far from solved yet. I hope our work can lead to new enthusiasm in the community and help the wide adoption of this technology in the near future."

Daniele De Martini, Postdoctoral Research Assistant at the ORI and Junior Research Fellow at Kellogg College

Towards identifying and closing gaps in assurance of autonomous road vehicles (TIGARS)

How can we improve the assurance of autonomous vehicles by adapting current engineering processes and technical analysis of first-generation autonomous systems?

TIGARS was an international research project which undertook research into techniques and engineering processes to address the assurance gaps and challenges for first-generation autonomous systems.

The project researched resilience and safety requirements, open

systems dependability, validation and verification of machine learning, security informed safety, and defence in depth and diversity. It conducted experimental trials with real-life demonstrator systems, and through numerous regulatory workshops in the UK and Japan the project furthered work in the international standardisation area.

The project concluded that an assurance approach based on argument-based mechanisms, specifically the Claims, Arguments, and Evidence (CAE) framework, will enable us to explore new mechanisms that could lead to advancements in safety assurance.

The team produced several TIGARS Topic Notes to support the development and evaluation of autonomous vehicles: www. adelard.com/tigars. These will be published by the UK Centre for the Protection of National Infrastructure as part of its guidance to industry.

Since the project concluded the project partners have been further developing and exploiting the results of TIGARS. This includes Witz and Nagoya University working on mobility as a service projects, that are taking place in several Japanese towns to demonstrate the value of safety assurance, building on the results of TIGARS.

Kanagawa University continues to be closely involved in standardisation activities. The standardisation work of the project has continued to support revisions of ISO/IEC/ IEEE 15026-2 Assurance case, ISO/IEC/IEEE 15026-4 Assurance in the life cycle and IEC 60300-1 Dependability management – Part 1: Guidance for management and application.

Adelard has been developing guidance and templates for safety cases for autonomous systems in a recent project and is also working for clients on the evaluation of security products and services that utilise AI/ML, building on the TIGARS work.

Project team

- Adelard LLP
- City, University of London
- Kanagawa University
- Nagoya University
- Witz Corporation

The TIGARS Experimental Vehicle



Automatic Testing Mechanism (ATM)

Towards an NCAP-like rating for robotics and autonomous systems

This project is designing a mechanism to automatically compute a safety rating, akin to an NCAP rating, for robotics and autonomous systems (RAS). This rating system will help users and regulators to identify and compare the safety of different RAS.

The system developed includes normal and adversarial trajectories for humans, the learned behaviour of the RAS, a policy on good human strategy around a RAS (or real human data), and a policy where the human would crash into the RAS. This all feeds into a distance computation and then a safety rating.

Preliminary evaluation of the testing mechanism and software framework on a high-fidelity autonomous car simulator has started.

Project partner

• Australian National University

This project is an exciting prospect for regulators and assessors of RAS in any domain. Being able to use the proposed testing mechanism and software to rate the safety of a system in an easily understandable way, would enable us to standardise the approval of such technology."

Professor John McDermid, AAIP Director and Non-Executive Director of HSE.





Safety Assurance of Autonomous Intravenous Medication Management Systems (SAM)

What if we could improve patient outcomes by using automated systems to administer just the right amount of medication at the right time to patients in intensive care units?

This project explored and addressed safety assurance challenges of robotic and autonomous systems (RAS) in healthcare using the example of intravenous (IV) medication management systems within an intensive care unit.

The focus of the study was the clinical system rather than the technology as such, and it looked at safety assurance challenges at the

The work on human factors in the design and use of

world use case to focus The Chartered Institute of Ergonomics and

Human Factors (CIEHF) contribution in the area of digital health

community of ergonomists, clinicians, technology developers and

regulators. They are continuing to build national and international

partnerships to provide guidance on current issues such as the role

and artificial intelligence. We have now established an active

autonomous infusion pumps provided a concrete, real-

intersection of engineering and human factors.

Four use scenarios at different levels of automation and autonomy were identified. These were used to explore stakeholder perceptions about risk, handover, and the investigation of adverse events involving autonomous infusion devices.

Three complementary analysis approaches Analysis Method, Systematic Human Error Reduction and Prediction Approach, NHS Digital SMART approach) were used to explore the safety issues around the use of autonomous infusion technology in intensive care.

(Functional Resonance

The project has made six recommendations aimed at technology developers, healthcare providers and regulators:

- Developers should consider patient experience and the impact on the patientclinician relationship.
- 2. Adoption of RAS should be accompanied by training to enable clinicians to maintain core clinical skills, and to educate clinicians about limitations of AI.
- Healthcare providers should consider the introduction of new AI specialist roles.



The intensive care set-up (photo credit: Dr Nick Reynolds, Royal Derby Hospital)

It is only natural that patients and their relatives have concerns where autonomous devices are used in a medical setting. I hope that patient involvement in the project helped to allay these concerns, because there is a real and vital need for this technology in the treatment of patients in ICU."

Howard Grundy, patient's relative

- Hazard analysis should be performed at the level of the clinical pathway or clinical system.
- Developers should design for situation awareness, handover between clinicians and RAS, and human performance variability.
- Regulators should promote existing best practices and establish an integrated safety governance framework for AI regulation in healthcare.

During the course of the project, the team established collaborations and partnerships with a number of bodies, including Chartered Institute of Ergonomics and Human Factors, NHSX and BSI, where these recommendations will be considered further.

Project team

- Human Reliability Associates Ltd
- NHS Digital
- University Hospitals of Derby and Burton NHS Foundation Trust

Dr Noorzaman Rashid, CEO Chartered Institute of Ergonomics and Human Factors

of AI in addressing global health inequalities."

Safety of the AI Clinician

Moving on from prediction: the safe use of AI in making medical decisions about sepsis treatment.

The next step in the use of AI in healthcare is an AIbased decision support system (DSS) that can suggest medication doses, supporting a clinician to make a decision about medical care. The team at Imperial College London was the first to develop an algorithm (the AI Clinician) that provides suggested doses of intravenous fluids and vasopressors in sepsis.

This project is investigating how to assure the safety of the AI Clinician for sepsis treatment in intensive care. Through this, it will help to establish general regulatory requirements for AI-based DSS. The project is structured around three key objectives:

- Review regulatory requirements in the UK and the USA - this will involve assessing the status of the current regulatory framework and any potential gaps, specifically for AI-based DSSs in healthcare.
- Define the required behaviour of the AI-based DSS for sepsis treatment – this will cover performance and safety, how to translate these requirements into its design, and how to detect and mitigate deviations of the DSS from the required behaviour.
- **3.** Deploy and test the DSS in pre-clinical safe settings to gather evidence on the safety and assurance.

Project team

- Imperial College London
- University of York
- NHS Digital

We are delighted to be working with Imperial College London and NHS Digital on this project. By bringing together our AMLAS process with Imperial's AI Clinician and NHS Digital's representation of the assessment authorities, we will be able to generate and evaluate a safety case that we hope will help move the UK towards safe deployment of AI in the NHS."

Dr Ibrahim Habli, AAIP

Assistive robots in healthcare

How can robots and artificial intelligence be used safely to improve the quality of life and increase independent living in an ageing population?

This project is investigating and evaluating the safety and regulatory requirements of close human-robot interaction (HRI) in domestic environments. It uses an overhead robotic system and the work is based around a series of experiments designed to validate a range of practical use-cases.

In combination with Environmental Survey Hazard Analysis (ESHA) the outcomes from workshops with stakeholders are being transformed into a set of functional and safety requirements. The next stage is to incorporate the use cases and scenarios into a detailed hazard analysis and design realistic experiments to assess the sensing and safety control requirements of the demonstrator platform.

The team has been using the data collected from the robot to develop the machine learning framework. They have developed some preliminary models which can detect the different transition



A compilation of the data from the sensors being used in the experiments with the CHIRON robot

states during a sit-to-stand activity and also detect the transition patterns of two different people.

Project team

• Bristol Robotics Laboratory, University of West of England

• Designability Charity Ltd

The project has provided essential support to the ongoing development of Environmental Survey Hazard Analysis. In addition to the experiments with students and project team members, the project has supported much of the development of the ESHA Foundational Ontology, which is the first stage in a radical improvement in the quality of the supporting evidence to safety cases that ESHA is intended to provide. The Foundational Ontology is general to any domain and will serve as the basis for developing more specialised ontologies. In particular, we are focused on the automotive domain (CAVs, driverless vehicles) and assistive robotics for the healthcare and social care domain."

Dr Chris Harper, Expert on System and Software Safety Assurance, Research Fellow on AAIP Assistive Robotics in Healthcare demonstrator, Bristol Robotics Laboratory, UWE





The PI, Praminda Caleb-Solly, has recruited a multidisciplinary team to develop a set of functional and non-functional requirements for physically assistive robots to provide mobility support. This work was followed up with a detailed use case analysis, using data from focus groups with domain experts in health and social care, and interviews with occupational therapists. These data have been carefully collected and analysed and provide a good foundation for future studies exploring physical human-robot interactions. I am very pleased to see the progress that has been made."

Professor Nigel Harris, Director of Innovation and Growth, West of England Academic Health Science Network. CEO Designability Charity 2009-2019

Human Factors in the Design and Use of Artificial Intelligence in Healthcare (HF/AI)

Out of isolation: developing guidance for regulators and technology developers that considers how an AI healthcare product will be used as part of the clinical system.

The aim of this new small-scale project is to provide guidance to regulatory bodies and technology developers on human factors (HF) in the design and use of AI applications in healthcare settings.

The project consists of two parallel and interacting work streams:

- Stakeholder engagement key stakeholders from regulatory bodies will support three stages of work: (i) identification of user needs with respect to HF guidance; (ii) providing feedback on guidance; and (iii) dissemination of the guidance.
- Guidance development the guidance will be shaped and developed through participation in the Chartered Institute of Ergonomics and Human Factors (CIEHF) special interest group on Digital Health & AI and existing literature on HF.

The project team has started establishing contact with a range of stakeholders through interviews and participation in events. An initial structure for the guidance has been developed and work on this will continue alongside ongoing collaboration with stakeholders.

Project team

• Human Factors Everywhere Ltd





CSI: Cobot industrial case study robot cell and a representation of this in the digital twinning environment

CSI: Cobot

Removing the cage and curtains: how can we assure the safety of cobots to support increased productivity in manufacturing?

The team began work by conducting a series of interviews with industry stakeholders to understand attitudes and concerns relating to robot safety. These helped to shape the direction of research and ensure the applicability of the designed safety systems. From this they have also developed a methodology to support the co-design of safe cobot processes, and increase knowledge of safety systems

amongst stakeholders. The card-based activity helps to engage stakeholders in the design process whilst also building confidence in the systems being developed.

The team have also constructed a lab-based replica of an industrial cobot manufacturing cell, enabling them to locally deploy and test sensors, and train machine learning algorithms to detect and track people and robots

working in the cell. In parallel, work has progressed on creating a digital twinning environment, including a representation of the physical cell, to provide a rich framework for analysing and controlling cobot processes both on- and off-line.

This has been extensively developed in response to Covid-19 (and in collaboration with other research projects) when access to physical

The impact of Covid-19 on the ability to access facilities has led us to develop a sophisticated digital twinning environment to support the development and testing of research methods, and facilitate their integration with real hardware. This has raised interesting questions about the extent to which digital twins can be used for assuring the safety of physical systems, but has also provided a framework to support training and remote operation of equipment – something our partners have raised as critically important in light of recent requirements around social distancing, both in terms of maintaining productivity and reducing risks to the health of staff."

Dr James Law, Acting Director and Director of Innovation and Knowledge Exchange, **Sheffield Robotics**

systems has been restricted. The framework is enabling the team to develop, integrate, and test techniques digitally, and explore the extent to which assurance can be carried out on virtual models of systems.

Particular safety and security advances have been made in the areas of hazard analysis and visual sensing to identify potential hazards and safety issues before they occur; in security policies and threat modelling to understand and mitigate cyber security risks; and stochastic modelling, including the development of a preliminary safety controller which can robustly respond to, and recover from, potential safety issues.

The next stage of the project will focus on integrating the safety methods with the digital twinning environment, and conducting initial testing.

Project partners

- University of Sheffield
- University of York

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

Saving lives, improving ocean data collection, and supporting green energy – how we can all benefit from *safe unmanned marine systems.*

Without suitable regulatory guidelines, Marine Autonomous Systems (MAS) are currently piloted and monitored by experts, keeping operational costs high and limiting the scale of simultaneously deployed fleets. This project will increase the safety of MAS by helping the vehicles identify the cause of any adverse behaviours.

The ALADDIN team is developing a monitoring and classification tool that correctly detects and diagnoses unexpected vehicle behaviour due to system malfunctions or environmental disturbances. The system will be validated using the case study of autonomous underwater gliders.

In the absence of accepted international terminology, the team has introduced new vocabulary for defining healthy and irregular behaviour of MAS. The terminology is designed to be transferable to any Robotic and Autonomous System (RAS) technology, although failure modes are specific to each system. A procedure has been designed to autonomously label time-series data as either standard or anomalous, with human feedback used to correct and improve tagging performance.

In addition, a range of MAS deployment datasets have been selected for use in the remainder of the project, enabling the design of transferable solutions. Remote-sensing data will complement the datasets to capture the environmental disturbances during the deployments.

Project partners

- University College London
- National Oceanography Centre

At present, as a result of limited situational awareness, it can be very challenging for highly trained glider pilots to detect sudden events such as wing-loss. The monitoring and classification tool developed in ALADDIN will enable us to detect, understand and react to even subtle divergences from expected vehicle behaviour, giving us the opportunity to implement mitigation to safeguard the vehicle and its valuable scientific data cargo."

Dr Catherine Harris, National Oceanography Centre

Safety of reconfigurable collaborative robots for flexible manufacturing systems (RECOLL)

Improving safety and productivity in manufacturing *by studying human-robot* interaction when using collaborative robots.

This project studied the safety-related human-robot behaviours (movements, layout occupation, voluntary/accidental contacts, near misses, etc) in a prototype machining production setup. They worked with the machining department of a manufacturing company (CEMBRE) in Northern Italy that extensively uses robot applications and flexible manufacturing systems supplied by MCM.

RECOLL was particularly interested in humanrobot behaviour when the operation needs frequent reallocation of human/ robot tasks, a changeable position of the human worker, and potential physical interaction with the machines.

The team developed a method for the quantitative evaluation of the safety risk assessment procedures, in line with current standards. This methodology can be applied to a wide range of human-robot collaboration scenarios. It allows us to understand if the preliminary risk assessment correctly includes all the possible hazardous situations and if its conservative approach will affect the cell's productivity.

Project team

- Machining Centers Manufacturing (MCM) SpA
- University of York
- National Research Council of Italy, Institute for Intelligent Industrial Systems and Technologies for Advanced Manufacturing

Demonstrator projects 31

Safe Airframe Inspection using Multiple UAVs (SAFEMUV)

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

Assuring the safety of teams of autonomous UAVs that collaboratively carry out safety-critical inspection tasks is hugely challenging. This project will develop a process for the systematic robustness assessment of UAV teams. This will be supported by the creation of tools for the specification, generation and testing of multi-UAV systems in simulation, lab-based settings and in the real-world.

The SAFEMUV project has carried out a comprehensive analysis of the stateof-practice in the assurance processes for UAV-based applications. Based on insights derived from this analysis, and meetings with regulatory bodies and their industrial partners, the team has developed the concept of operations needed for delivering the safety analysis process.

The team has also started to implement a prototype pipeline to support the definition of multi-UAV system and scenario specifications from stakeholders. These specifications will be transformed into executable models for the simulator of the SAFEMUV demonstrator as well as the code required for the real multi-UAV application.

The next step involves the realisation of the safety analysis process for the SAFEMUV demonstrator. Lessons learnt will enhance the prototypical pipeline for the specification of collaborative multi-UAV inspection scenarios with key quality criteria that will support the analysis of these scenarios in simulation, lab-based settings and in the real-world. The developed pipeline will be generalisable to other application domains where robotic teams are employed.

Project team

- University of York
- SnT, University of Luxembourg
- Bonn-Rhein-Sieg University, Germany

SAFEMUV comprises an international team of experts in safety assurance, robotic systems and system verification from academia, industry and regulatory bodies. Working as a team has enabled us to produce new knowledge and make more progress than each partner could have achieved in isolation."

Simos Gerasimou, Lecturer in Computer Science, University of York



Mediumsized AGV for soft-fruit production (MeSAPro)

Using autonomous robotic systems to support human fruit pickers and reduce workplace accidents.

Researchers from Saga Robotics and the University of Lincoln have developed fleets of autonomous soft-fruit production robots that undertake tasks including monitoring, forecasting, harvesting and transportation in real-world fruit farm environments.

This project will complement existing work on developing the functionality and technical capabilities of the robotic system, by focusing on safety assurance of the soft-fruit production RAS:

- defining safety requirements for the senseunderstand-decide-act components of the RAS
- developing methods to detect deviations from safe behaviour and ways to mitigate the effect of such deviations
- formally verifying the sensing, understanding and deciding components of the existing software stack for the RAS

With the help of industrial partner Berry Gardens Growers, the team is developing several harvesting scenarios, which will help them to specify the interactions between humans and RAS.

Project team

- Lincoln Centre for Autonomous Systems, University of Lincoln
- Saga Robotics



Safety Assurance of Cooperating Construction Equipment in Semi-Automated Sites (SUCCESS)

Moving towards the use of a fleet of autonomous construction vehicles by assuring their safety when they work together as a system-of-systems.

The SUCCESS demonstrator project explored numerous aspects in the safety assurance of cooperating systems-of-systems (SoS), with a special focus on the construction machinery domain.

Specifically, the project:

- highlighted the challenges in hazard analysis of SoS
- demonstrated a digital twin-based approach for simulation and verification of safety requirements
- evidenced the potential of geofences for dynamic risk management
- outlined an end-end tool framework for safety analysis

The project used a quarry site at Volvo Construction Equipment - the "Electric Site" - to give a real environment in which to study the fleet of machines working together as a SoS.

The team proposed new ideas for the hazard analysis of SoS, including

The ambition of this project was to run a fleet of autonomous vehicles in a real environment to identify ways to assure the safety of this complex SoS. The results and lessons learned on SUCCESS will be utilised through my work at Volvo Autonomous Solutions." Stephan Baumgart, Pilot Safety Manager, Volvo Autonomous Solutions

a state-based extension to System Theoretic Process Analysis (STPA) and the team's "SafeSoS" approach. SafeSoS offers a hierarchical process for specifying a SoS and using those specifications for performing a safety analysis.

The Volvo training simulators at Mälardalen University were adapted and extended to form a digital twin of the Electric Site. This enabled the team to verify the safety requirements of their collaborating autonomous systems and to implement and check multiple safetycritical scenarios.

In addition, the project team has shown the potential of different types of geofences for dynamic risk management during the operational phase. They also defined an endto-end tool framework that incorporates both process and product-based evidences, safety contracts and dynamic assurance.

As an independent assessment body undertaking third party assessments we have been able to use the SUCCESS project's work to determine what is feasible within existing regulations and safety standards and, when applicable, what could be used to support the development of future standards."

Henrik Thane, CEO, Safety Integrity AB

Read more about the project and the guidance developed: bit.ly/aaipsuccess

Project partners

- Mälardalen University, Sweden
- Volvo Construction
 Equipment
- Safety Integrity AB

The Future

In 2021, we will further develop the practical guidance we offer by creating a standalone website for industry and regulators, to provide a useable resource for the assurance and regulation of autonomous systems.

his resource builds on the work described earlier (see pages 12 and 13) on the five building blocks of safety assurance: assurance of machine learning, safety assurance of understanding in autonomous systems, safety assurance of decision making in autonomous systems, safety assurance of autonomous systems in complex environments, and social acceptability of autonomous systems. It will provide an accessible process for using the guidance developed in each of these key areas.

Our work on ethics will also be strengthened over the coming year. In 2020 we spoke with members of the public about their views on autonomous systems. One focus group suggested such technology needs to be "acceptably safe" and that's what our work will consider: what does acceptably safe mean to the different stakeholders of autonomous systems? How can we ensure the system meets such requirements and how can we communicate this to those that need to know? We are also looking forward to hosting the 40th annual SAFECOMP conference, with a special theme in 2021 of safe human-robotic and autonomous system (RAS) interaction. Whether in person in York, or online, this is an important date in the diary to exchange insights and experience on emerging methods, approaches and practical solutions.

We will also put a physical stamp on our work with the opening of the brand new facility for the Institute for Safe Autonomy. This new building, which AAIP and our funder Lloyd's Register Foundation have been fundamental in establishing, will provide specialist laboratories and testing facilities, to support the Institute's work to ensure that robotics, their software, and their communications are safe.

By bringing together AAIP and complementary University expertise with industry and other stakeholders in one facility, we will strengthen our capacity to advance the safety assurance of autonomous systems. The year will bring our existing work on standards and regulations into sharper focus as we look to fund more demonstrator projects led by regulatory bodies. Building on our existing work with regulatory organisations, and using the practical guidance we have already developed, we want to move the community towards flexible standards and regulations that work for ever-evolving autonomous technologies.

As ever, the technology continues to advance and mature. Through our collaborations across the globe, our research and guidance are keeping pace and in 2021 this will help us to work with the community to pioneer the new global safety standards that will keep us all safe.

Work with us

The challenge of assuring the safety of RAS requires further funding, collaborative research, technology case studies, evolving regulations, and more. Contact us if you would like to collaborate with us and 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
- www.linkedin.com/company/assuring-autonomy
- https://medium.com/assuringautonomy.medium.com/

January 2021 AAIP ethics workshop: Ethically-informed engineering of autonomous systems of autonomy

0

April 2021

Panel discussion at Algo 2021 Conference: Building Ecosystems of Excellence & Trust

Summer 2021 Digital health safety conference with NHS digital

August

September

October

January

February

March

April

Мау

June

July

January 2021

0

Trustworthy Autonomous Systems Node in Resilience Partner Workshop

February 2021

Assurance of Machine Learning for Use in Autonomous Systems (AMLAS) handbook published

February 2021 Safety-Critical Systems Symposium

 \bigcirc

June 2021

0

Safer Autonomous Systems Workshop at DSN 2021

October 2021 Institute for Safe Autonomy opens



