The Assuring Autonomy International Programme 2018-2022

A Review of Emerging Outcomes

By Independent Consultants

Paul Rhodes and Alan Graver

January 2023

Version: Final

Table of contents

Acknowledgement

The consultants would like to thank Sarah Heathwood, Professor John McDermid, and Dr Ana MacIntosh for all of their insights, openness and support during this review.

1 The Assuring Autonomy Programme (AAIP)

For the University of York, and Professor John McDermid[i], the origins of the research excellence approach for AAIP began in the 1980s, when they began to establish themselves as an international leader in safety of complex computer-controlled systems and software.

In October 2016 Lloyd's Register Foundation (LRF) published its 'Foresight review of robotics and autonomous systems Serving a safer world'[ii] in which it concluded that:

> *"There are some important areas which need addressing if we are to see the safety benefits from the implementation of RAS, and where the Foundation may be well positioned to lead or support other international efforts."*

The report goes on to suggest priority areas for focus in future linked to openness and sharing, security and resilience, public trust, understanding and skills and **assurance and certification** – aligned to which are recommendations around asset self certification and assurance of Robotics and Autonomous Systems (RAS) learning systems.

> *"These recommendations are mostly in the area of the assurance of autonomous systems. This is because there is 'white space' where the Foundation can bring additional value. However once progress is made in the assurance of safe autonomy, there will be knock-on implications for the future design and build of the software and hardware aspect of RAS."*

Reflecting on the progress being made by the AAIP Programme at the November 2022 collaboration workshop, six years on from the LRF Foresight Report, John McDermid OBE FREng neatly summarised

> *"AAIP has been trying to colour those white spaces in."*

LRF funded the Assuring Autonomy International Programme (AAIP) to address the assurance and regulatory challenges identified in the Foresight Report and more widely support the safe introduction of RAS to realise their benefits for society.

The initial funding award was for £10m over five years, from 2018-2023. The University of York also contributed £2m.

The original funding bid summarised the 'need' for the Programme.

> "The Need: Current analysis methods and regulatory frameworks do not fully cover the technologies, especially AI, used by RAS. Furthermore the groups developing and deploying RAS are not always familiar with standard safety, assurance and regulation practices. There is therefore an urgent need for improvement in methods for assurance and regulation of RAS – so these processes can catch up with, and positively influence, RAS developments and operations and society can benefit from the safe use of RAS."

The Assuring Autonomy International Programme, as funded by LRF, would make a significant step towards addressing these challenges by:

- Creating an International Community of Practice including industry, academia and the regulatory community
- Undertaking a series of major translational research activities intended to influence industrial practice through collaboration with ongoing and new real-world 'demonstrator' projects from which key safety lessons on the deployment of RAS will be drawn
- Leading an international collaborative activity to develop a suite of cutting-edge professional education and training programmes, making them available to industrial and academic partners worldwide

Specific technical aims will include:

- Establishing agreed means of assuring autonomous systems
- Evolving design practices so that systems are 'risk aware' and can self-assess and self-certify.

Exploration and discovery

> *"What is complex has changed over time, which is what continues to make this field so exciting." (Professor John McDermid, AAIP collaboration workshop November 2022)*

Since 2018, the AAIP has been trying to address some of these questions through its blended research and demonstrator projects approach:

- Why is assuring autonomy, using AI and ML, tricky?
- What is different with autonomy?
- How to address the AS safety assurance delta with confidence, consensus and acceptance?
- How do we do safety analysis?
- What does acceptably safe mean for the system?
- Is 'it' (AS / AI / ML) safe?
- How do move to trials (e.g. in a healthcare clinical context)?
- What does ethically safe / acceptable mean prior to the deployment of an AS?
- How can we manage the legal issues in autonomy?
- How can a model of the world be built by AS with an appropriate fidelity and accuracy?
- How to do safe RAS DevOps in future?
- What is the difference between a DSA safety case and one using HAZOP?
- How do we safely assure AS in complex environments?

> *"We have done things that are based on principles, but also that are practical to help people with real-world problems." (Dr Richard Hawkins, Senior Research Fellow, AAIP team, collaboration workshop November 2022)*

Intended outcomes

A logic model was produced at the start of the AAIP describing the intended objectives, expected outputs, outcomes along with indicators (measurements of change) and desired longer-term impact (which would be unlikely to happen within the five-year lifetime of the Programme rather emerge beyond that).[iii]

Appendix A: Research Impact and Outcomes

Table 2: Objectives, Outputs, Outcomes, Outcome Indicators and Impact

Objectives – The aims of the project	Outputs – A project's activities, services & products	Outcomes – What will change?	Outcome indicators – How will you measure the change?	Impact – What is the ultimate benefit to society in line with the Foundation's charitable purpose?
· Establish agreed means to assuring or certifying autonomous and adaptive systems, as common as possible across domains · Evolve design practices so systems are "risk aware" and can self-certify · Improve education and training for safety specialists and systems designers to ensure safety of autonomous systems · Enable education and training for people "displaced" by autonomy	· Analysis and certification frameworks for systems using artificial intelligence · Design principles for systems using learning and artificial intelligence · Dynamic risk assessment frameworks for systems that learn in operation · MSc and CPD courses available (potentially franchised) for safety engineers and developers of autonomous systems	· Increased ability to certify systems that use adaptive techniques and artificial intelligence · Increased ability to "design for assurance" · Increased proportion of systems are "risk aware" and self-certifying · Greater alignment of skills with needs in industry and regulators · Greater acceptance of autonomous systems, as operational staff are able to move to other jobs	· Ease and speed with which autonomous systems are approved for operational use · Increasing number of systems carrying run-time (safety) certificates · Use of qualifications from MScs and CPDs as criteria for recruiting and training staff · Greater take-up of focused retraining by operators affected by autonomy (reduced resistance from unions) · Rational debate on proposed systems,	· Greater deployment of autonomous systems that net reduce risks to society and infrastructure · Systems detect their own problems effectively, reducing operational risk · More effective and cost-effective systems, due to increased skills of staff working on autonomous systems · Possible job losses from use of autonomous systems is significantly reduced · Society benefits from appropriate use of autonomous systems

It is challenging to directly track the progress clearly being made by the AIIP when compared to the outcomes and indicators as identified in this original logic model rather:

- There have been some valuable, short- and intermediary-term outcomes observed through the review that could conceivably develop into the outcomes originally laid out. This means that it may be taking longer for the outcomes to emerge than originally anticipated (for example as the 'landscape' for assuring safety was less developed than first thought) and / or that the outcomes emerging have iterated from the original intentions
- There have been some outcomes observed for different stakeholder audiences and these are presented in the conclusions.

In light of this independent observation the AAIP may wish to revisit the logic model and revise it accordingly with past performance in view, whilst also orienting it towards the ambitions for the next chapter of the AAIP linked to its five research pillars reflecting a safety and assurance process for RAS

1. Societal Acceptability of Autonomous Systems (SOCA)
2. Safety 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. Assurance of Machine Learning for Autonomous Systems (AMLAS)

There have been a number of achievements:

- ✓ The AAIP is currently producing domain-agnostic manuals/guides for each pillar which will link to form a coherent whole and are adaptable across domains.
- ✓ 'Assurance of Machine Learning for use in Autonomous Systems' (AMLAS)[iv] has been downloaded over 1,000 times by people in 18 countries and 20 sectors. Over 500 people have been trained at AMLAS workshops since 2021
- ✓ The only MSc module dedicated to the safety assurance of RAS has been developed and is being delivered
- ✓ Over 50 health professionals have been trained via bespoke CPD
- ✓ The AAIP team has exerted influence with Government (DSTL and DfT) on Artificial Intelligence (AI), Machine Learning (ML) and autonomy.
- ✓ The team are overseeing BSI work on Autonomous Vehicles (AV) standards on assuring ML.

Remarking on the overall achievements of the AAIP relative to the scale of the challenge Professor John McDermid said in November 2022:

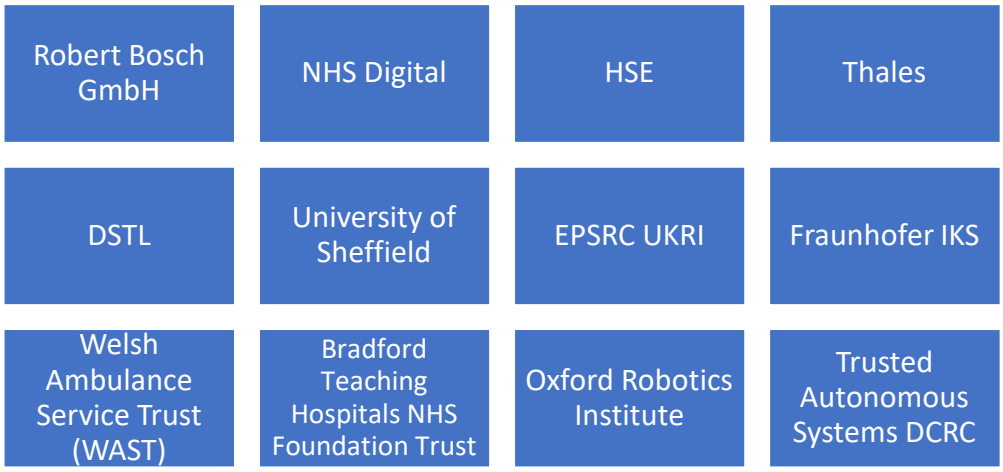
> *"The investment of £10m is a valuable contribution to what is closer to a £1bn problem, as there was a hope that AAIP would work across all technologies, all domains and jurisdictions.*
>
> *We've done pretty well I think, and the leverage has been good, plus the dedicated building we now have could provide an environment for work and testing we didn't have before."*

2 Methodology

This review of emerging impact was carried out by independent evaluators, Paul Rhodes (Paul Rhodes Consulting) and Alan Graver (Skyblue Research Ltd), between September and November 2022.

The following approaches were employed:

- 13 depth, client-selected case studies[v] completed and analysed
- Learning lessons reviews with the AAIP Programme team
- Review of the research that led to the Programme, the initial bid and logic model and annual reviews 2018-2021
- Rapid review of demonstrator project aims and publications
- Attendance at AAIP Collaboration Workshop (Day 1, November 2022)



Limitations

Managing the reader's expectations

Whilst the consultants recognise that the AAIP has been a process of research excellence applied to autonomous systems, this report is not a review of the quality of the research undertaken or the efficacy of the outputs generated. That is better achieved through the peer review and feedback processes that have been involved in bringing the fundamental research (papers and publications), body of knowledge, demonstrator projects, education and training and public engagement activity to life.

However, this report does reference citations of work produced and encourages the reader to explore the wide range of research papers, guidance and studies or articles published and peer-reviewed – and archived – since the AAIP commenced for a much fuller appreciation of the depth and breadth of research and practice that has been undertaken and shared for the benefit of all.

3 Key Findings

3.1: AAIP's contribution to enhancing design and assurance practices

How are industrial practices safer because of the AAIP's work?

The challenge

Use of Robotics and Autonomous Systems (RAS) is already having and will have a significant impact on safety and quality of life for society. New technology brings new risks. Assuring and regulating safety of RAS is currently the biggest obstacle to gaining the benefits of these new technologies[vi].

> *'Safety must be 'by design,' but the complexity of interactions between multiple RAS systems means that ensuring system safety will depend on standards and clear understanding of the impacts of modularisation.'*[vii]

In 2017, the AAIP recognised that prevailing analysis methods and regulatory frameworks did not fully cover the technologies, especially Artificial Intelligence (AI), used by RAS. Furthermore, the groups developing and deploying RAS were not always familiar with standard safety, assurance and regulation practices. Moreover, this need is exacerbated by the requirement to design for uncertainty and design for safety.

> *'Safety starts with the design process and the design process starts with an analysis of function. Designing a safe system is easier when the interaction between the system and its operating environment is constrained and well defined.'* [viii]

Safety assurance practices across industry were thought to be inconsistent and lagging behind RAS technological developments.

> *"Sometimes we forget how big the gap is around understanding. Nobody else has articulated how you would go about dealing with assessing ML in such a way you could integrate that assessment into a classical safety process / safety case. We use the term assurance, but the focus is really on assuring safety." (The AAIP Programme, reflections in September 2022)*

As one of the case study (Dstl) interviewees summarised:

> *"There is plenty of expertise in assuring traditional platforms but there is nothing like the maturity in evidence when it comes to developing assurance arguments to achieve certification for autonomous platforms. There is much more work to do and ultimately, 'assurance is the key that safely unlocks the potential of artificial intelligence and autonomous systems'."*

In this context the AAIP has encouraged collaborators to enhance their design and assurance practices. One of the most tangible assets that can contribute towards this ambition, is the production of a world-first leading methodology, presented as systematic and structured guidance known as **'Assurance of Machine Learning for use in Autonomous Systems' (AMLAS).**[ix] AMLAS comprises a set of safety case patterns and a process for

systematically integrating safety assurance into the development of machine learnt (ML) components. This provides a compelling argument for any ML model to feed into the wider system safety case.

> *"AMLAS is our process for assuring ML for autonomous systems and can help demonstrate to people it is safe to use. We aimed to provide not high-level objectives, rather something really practical if someone wanted to develop a safety case. We've described the process, the expected activities you should follow, the artefacts that should be generated but also a set of safety argument patterns – and how you use the 'stuff' generated to develop your safety case.*
>
> *The AMLAS tool can help you build safety arguments too, but we want to develop that into a proper tool for industry to use. We think AMLAS is fine for relatively small-scale examples, but by January 2023 we want to work out how to build more effective 'tooling' support for these processes, to assist projects at scale and to help manage change. Over time we will refine AMLAS guidance" (Dr Richard Hawkins, AAIP team, collaboration workshop November 2022)*

AMLAS was the first of a series of guidance documents since produced by the AAIP. Guidance was produced by demonstrator projects helpfully contributing to the ever-growing Body of Knowledge (BoK).

The approach and presentation of AMLAS was so favourable reports the AAIP team that it largely inspired the approach and aesthetic developed for the Safety Assurance of autonomous systems in Complex Environments (SACE)[x] guidance published in 2022.

A dedicated guidance website has been developed including proof of concept tool for AMLAS, which will iterate to become even more industry- and user-friendly in future.[xi]

The difference being made

> *"We have built models and life-cycles to help assure machine learning e.g., SUDA (Sense-Understand-Decide-Act), and our guidance has been built on these models and cycles."*

Demand for the AMLAS guidance is evident from the interest expressed by the international community since its publication in February 2021. There have been over 1,000 downloads from individuals working in 20 sectors / disciplines / specialist areas of interest and from institutions or organisations located in 18 countries.

Whilst further work is required to better understand the extent to which this guidance is being used in everyday practice by those that have downloaded it, insights have been collected through a series of depth case studies and sampled survey returns in 2022. Results suggest that the process of shaping the guidance, using it directly or raising awareness of the structured approach amongst others in their organisations and sphere of influence has been extremely positive, **saving them time, money and invention**.

The insights suggest that:

✓ AMLAS guidance is valued for both its structure and rigorous technical content
✓ Using AMLAS enables the conditions for the creation of stronger safety arguments
✓ Developed by 'domain agnostic' safety specialists, it invites cross-occupation conversations, and therefore different questions and greater critical challenge
✓ AMLAS provides a stronger basis for challenging and testing safety cases. It is not a 'plug in solution' but requires customisation to fit the context it is being applied in
✓ Project teams using AMLAS should result in a default approach that brings experts in safety and product design together, with the result that the safety case is more systematised and widely understood across multiple occupations and roles
✓ The AMLAS structure provides a common structure for teams to use, and benchmark against their bespoke safety cases
✓ Using the guidance encourages outward thinking and learning from other sectors
✓ It has provided guidance on developing safety cases for systems including elements of ML
✓ It has provided clear guidance on the implementation of algorithm on practical applications e.g., for algorithm design; case implementation; techniques dissemination (one example was how AMLAS has been used in marine autonomous systems for Remotely Operated Underwater Vehicles (ROVs)

These assertions are exemplified through the following case study extracts

Exhibit 1: Healthcare sector. Case study acknowledgement: Shakir Laher, AI/ML Research Associate / Safety Engineer, NHS Digital

Since starting at NHS Digital as a software engineer, Shakir is now a healthcare informatics specialist developing software and leading research projects in Artificial Intelligence / Machine Learning assurance focused on safety.

> *"NHS Digital has a remit to ensure technology is integrated, deployed and used safely to treat patients. AAIP has a strong focus on safety assurance producing high level domain agnostic guidance. Therefore, we began collaborating with AAIP to help us understand our gaps and limitations and to take relevant guidance and implement that in the healthcare domain to assess its applicability and give back to the state of the art. A specific guidance we have reviewed is AMLAS. It was primarily written for manufacturers of ML, so we have embarked on extending it for 'adopters[xii]' (e.g., clinicians) of ML so that they can assure safety of the technologies in their clinical pathways.*
>
> *This approach has been inspired by AMLAS and through working together. Consequently, we published a review paper[xiii] with a recommendation that AMLAS was fit for purpose as a safety assurance methodology when applied to healthcare ML technologies, although development of healthcare specific supplementary guidance would benefit those implementing the methodology. That supplementary guidance has since been created and will be published in late 2022."*

Asked if AMLAS has impacted safety working practices within NHS Digital, Shakir reflects that there have been a range of positive outcomes.

"It's consolidated the information, and misperceptions we had. It's provided a clearer path for what we need to do for the next few years and consolidated our thinking towards tangible outputs. It also spurred us on to more work, to extend AMLAS and include some extra stages. If AMLAS is going to make up part of the ML safety assurance process at national level, then manufacturers and adopters will be able to use that and our additional guidance with confidence."

Exhibit 2: Automotive sector. Case study acknowledgement: Lydia Gauerhof, Research Engineer for Robert Bosch GmbH

Since starting at Bosch in 2016, Lydia's work has been focused on safety assurance of artificial intelligence applied in automated driving. Lydia first began working with AAIP team members by contributing to the review of the draft AMLAS guidance. One of the fully worked examples Lydia helped to develop, on pedestrian requirements was included as part of AMLAS. Lydia has shared the AMLAS guidance across the business through facilitated workshops with the AAIP team and via electronic means. She works with embedded safety experts in R&D teams within business units to make them aware of the guidance and to convince them of its value, particularly from bringing academia and industry together:

"Universities focus more on methods; Bosch focuses more on products. Of course, methods are needed for development and here is where the value of universities comes in through their cross-domain expertise."

Lydia feels that AMLAS provides a resource and approach that can be shared across project teams. It can be used as a credible basis of discussions about assuring safety. The alternative is that a bespoke safety case is developed, which is typically then only understood by those who created it.

"AMLAS is important as it provides the know how to develop a safety argument. The structure really works; it is very clear for people who come from an AI rather than a safety background. There will be special details that are outside of the guidance which are domain specific. We can use the guidance as a sanity check, asking for example, 'did AMLAS use the same argument?'"

AMLAS has the potential to impact safety working practices within Bosch.

"AMLAS is not a 'plug in solution' but requires customisation to fit the context it is being applied in. The guidance could inform, and therefore improve the safety of, a wide range of functions in development by Bosch."

Through her corporate research role and influence she can support more holistic safety arguments by technical specialists and others in the business units, with greater consideration of the quality of data a design will generate about safety.

"AMLAS is helpful as there are a range of different perception functions. You have pedestrian detection, traffic sign recognition, semantic segmentation, and so on."

The recognition of the importance of having systems that can demonstrate both performance and safe performance has increased over the past two years.

> *"It is easy to do, but hard to do safely."*

Exhibit 3: Defence sector. Dstl (anonymised)

The AAIP is contributing to the defence sector's exploration of how to make autonomous and artificial intelligence systems do what they are supposed to do within the field of advanced and dependable autonomy.

> *"When AMLAS was first produced, we recognised that the AAIP was making great contributions and through our opportunity to peer review AMLAS, we knew the work would be of good quality. Whilst Dstl has produced a 'Biscuit Book'[xiv] which has a specific high-level purpose and target audience within the MoD and beyond, I especially like AMLAS because it is much deeper; the technical content has strength. It gives you in-depth sight of what you need to put in place a structured argument for machine learning assurance.*
>
> *Time is a challenge in our professional work, so it is a mark of esteem that we have elected to spend time working closely with the AAIP on peer-reviewing AMLAS; and on supporting the review of SACE[xv] - which we believe provides a useful system-level accompaniment to AMLAS, and we're absolutely delighted to see this development; and delivering joint presentations in influential settings amongst our international community such as the 'Five Eyes nations in The Technical Cooperation Program' (TTCP)[xvi]."*

Asked whether AMLAS has impacted working practices at Dstl our interviewee felt that this needed some reflection.

> *"As a research lab our autonomous programmes are not geared up for the level of assurance that AMLAS is designed for i.e., systems that are to be put into operational service. Our work is around 'what is it possible to achieve' and wouldn't necessarily require the expense of going through the AMLAS lifecycle. However, we are thinking about how we develop the assurance case for some systems such as larger autonomous vehicles …and this is where we think having AMLAS is going to be beneficial in future."*

An example of AMLAS being used in practice

At the AAIP collaboration workshop (November 2022), the team presented findings from a demonstrator project that applied AMLAS to a real-world example. This represents the first fully developed safety case for an ML component containing explicit argument and evidence as to the safety of the machine learning.'[xvii]

> *"AAIP helped Craft Prospect Ltd (lead of the ACTIONS demonstrator project) apply AMLAS to their neural network approach for wildfire detection technology. This has been done in simulation but will be applied to the real satellites in future."*

Wildfires are a common problem in many areas of the world with often catastrophic consequences. A number of systems have been created to provide early warnings of wildfires, including those that use satellite data to detect fires.

The increased availability of small satellites, such as CubeSats, allows the wildfire detection response time to be reduced by deploying constellations of multiple satellites over regions of interest. By using machine learnt components on-board the satellites, constraints which limit the amount of data that can be processed and sent back to ground stations can be overcome. There are hazards associated with wildfire alert systems, such as failing to detect the presence of a wildfire, or detecting a wildfire in the incorrect location. It is therefore necessary to be able to create a safety assurance case for the wildfire alert ML component that demonstrates it is sufficiently safe for use.

The paper describes in detail how a safety assurance case for an ML wildfire alert system is created. The AAIP team knows that it is early days to make claims about AMLAS' wider application in industry yet, however, further case studies like this are being created to build the practical evidence base over time.

> *"When we have more case studies we can generalise." (Professor Ibrahim Habli[xviii], AAIP team, collaboration workshop November 2022)*

Added value

NHS Digital

Shakir reflects:

> *"We might have got to a similar process to AMLAS based on existing literature relating to life cycle safety assurance, but we never had the dedicated resource to achieve it in the same timeframe. Where the credit is due to the AAIP is that they tell you 'how to do' not just 'what you need to do'.*
>
> *The expertise in York, the level of detail and depth is incomparable to anywhere else. If the AAIP wasn't there you would immediately miss the intelligence – the high-level interactions and intellectual conversations that can help you with your chain of thought and put you straight."*

Robert Bosch GmbH

Lydia reflects that collaboration may still have occurred in the absence of AAIP, however, it would have been more informal, less focussed and therefore less impactful. The added value of agreeing to become an AAIP Fellow has been access to safety expertise in a safe environment.

> *"The York team had a good understanding of machine learning components. It was firstly useful to have this sanity check that we had the same understanding, then we could extend our respective knowledge. We could also talk about similar problems we'd encountered, but in different domains. We could do this safely, and without sharing commercial confidences."*

Dstl

AAIP is not the only source that stakeholders in the defence sector will rely on, however the expertise appears to be valued and offer distinct benefits to those who know about its capabilities as seen from this comment by a representative in the defence sector.

> *"Whilst we do work with others in a similar way to AAIP for AMLAS, it is really helpful to have it there as without it there would have been something missing. The analogy is that we can see different practice across our partners, and AMLAS is more than a box of bits – it comes with instructions and an assembly guide to help you make that safety case argument in a controlled way. It is definitely more than safety argument templates too, it's like a handrail that helps you know how to go through producing it. It gives us an avenue to go forward and say, 'we have a mechanism, and we are confident we know what to do to produce a compelling safety argument'."*

AAIP recently presented some of its work at West Point to Five Eyes nation representatives in the defence sector. This was a workshop organised by US Navy and Army research labs and explored the application of AMLAS to a number of different AI projects.

Where next?

NHS Digital

Within NHS Digital Shakir has started a process of promoting AMLAS and supplementary guidance which will be hosted on the national NHS Digital website from 2023. This has the potential to be accessed by thousands of professionals. If these approaches are adopted by the newly established Integrated Care Boards (ICBs), potentially this will impact large numbers of clinical cohorts. AMLAS has also been explicitly referenced as part of the 'BS 30440 Validation framework for the use of AI within healthcare – Specification[xix]' due for publication in January 2023.

The next phase of collaboration with the AAIP will focus on further research into safety assurance of ML in systems:

"Specifically, we hope to evaluate the safety of image-based systems used for diagnostic predictions and learn more towards how we monitor their safety while in live operational use."

Robert Bosch GmbH

"I would really like to work on a joint publication and agree a speciality with the AAIP team. That would give our collaboration a deeper purpose."

When asked her thoughts for the next phase of the AAIP, Lydia was keen that AMLAS and other guidance are updated to reflect increased knowledge around machine learning, data selection and learning as a result of having vehicles in use.

"Active learning and iterative improvement using machine learning will be helpful. We have the structure, but now we need to understand how the safety case changes as the systems are used. As artefacts change, how does that impact on the wider system?"

Lydia felt that AMLAS could have greater influence at an automotive sector level if it was embedded in the new international standard for road vehicles – safety and artificial intelligence, ISO / AI PAS 8800.[xx]

The AAIP team reported in November 2022 that they have plans to develop the guidance and tools they have developed to be more widely used across industry but also to develop new methodologies that cater for AS deployment and the need to assure its safety during operation.

"Guidance has so far been produced prior to AS deployment, creating a safety case for an ML component in an AS. Some of our future work considers how we ensure a safety case for ML remains appropriate in operation i.e., post-deployment, looking at ways to minimise the rework that might be required. This is early days, and not easy. We're exploring this through the RAILS project." (Dr Richard Hawkins, AAIP team, collaboration workshop November 2022)

Conclusion

AMLAS can help individuals that work in a role that includes the assurance of safety to know what needs to be done differently in their organisation's practice to assure the safety of machine learning. With its adoption, at scale, over time, the AMLAS guidance has the potential to inform the improvement of safety practices across multiple domains across the world. Further plans to develop this, and other guidance and tools, should encourage the conditions whereby more industry partners use it in ways to develop robust safety cases, in turn building an evidence base for its practical application in real-world environments and contexts.

3.2: AAIP's contribution to validating design and assurance capabilities through translational demonstrators

How has the AAIP impacted safety-critical sectors?

The challenge

> *"The problem is that if you have people developing this technology but with no background in regulation there will be a deployment challenge (of RAS). The gap between really clever 'stuff' and demonstration to deploy and sustain that will not be injurious to people and environment." (The AAIP Programme team, reflections in September 2022)*

The LRF Foresight Report (2016) suggested that:

> *"There is [therefore] an important impact on the safety of people and of their environments. There is also a need to build RAS systems safely, so they act dependably and appropriately in all situations, including when they fail."*

Through this lens therefore the AAIP agreed that a key to delivering the Programme would be the use of demonstrators. These are real-life projects testing or deploying RAS 'in the wilds' that would be supported by the Programme both to develop and to evaluate approaches to assurance and regulation, including guidance.

> *"Safety is not a hard science it's really experiential." (The AAIP Programme team, reflections in September 2022)*

By ensuring a manageable, but purposely diverse, portfolio of demonstrators across domains and safety critical sectors, these experiences would reveal needs, challenges and learning. They would provide a focus and basis for improving and validating the Programme's emerging guidance (please see AMLAS referred to in Chapter 1).

The Programme has invested £5,041,370 in 24 demonstrators[xxi] comprising: eight in health and social care, four in automotive, three in maritime, three in manufacturing, two in aviation and one each in mining, space, agriculture and quarrying.

> *"We are leading and funding collaborative research projects across the globe to develop methods for the assurance and regulation of robotics and autonomous systems (RAS). Much of this research is shared through the Body of Knowledge.*
>
> *While our researchers are working in particular domains, the lessons and guidance that come from their research are often transferable to other sectors." (AAIP website)*

The reader is encouraged to review the detail about each demonstrator via the AAIP website and the annual reviews since 2018 that provide further insights about the experience of preparing and delivering each collaboration. This report highlights just a few examples of the difference being made through the investment in demonstrators from a sample of collaborators.

The difference being made

> *"We have connected with people that are working on prototypes, to work with them on the assurance and regulatory issues for the system they are working on. We tried to identify projects that will influence practice in future." (Rationale for demonstrator selection, AAIP team, collaboration workshop, November 2022)*

In the view of the Programme Team, the demonstrators have been a vital ingredient since 2018. They achieved different things individually, but also had a collective value whether they have been focused on building, developing or testing aspects of a system; validating assurance capabilities through the development of improved safety cases and / or in developing guidance or supplying learning that could be used, for example, in AMLAS.

> *"Earlier demonstrators developed guidance as the applications were all about where they could contribute to the Body of Knowledge. Later demonstrators have helped us validate our guidance, and the most recent commissions have been even more closely embedded to strongly align with the Programme's Research Strategy. This means a greater percentage of AAIP Team time set aside to work alongside each one. The point of them is that other people can learn from what they do." (The AAIP Programme, reflections in November 2022)*

Insights collected as part of this review suggest that the demonstrators have led to a number of benefits including, but not limited to:

✓ Changes in understanding about how to safely assure ML components / systems
✓ The strengthening of the approach to an assessment of compliance
✓ Shifts in attitudes to develop a positive, concomitant route for AI and safety assurance (by regulators)
✓ The influence of thinking that goes into the development of standards that embrace advances in technology from a much more informed (user-experienced) position
✓ A contribution to thinking that goes in to the development of policy on assurance of autonomous systems (by regulators)
✓ The creation of enduring assets (intellectual know-how, research papers, datasets) that can benefit a range of sectors
✓ The application and extension of methods that can be used for verification and testing across a range of safety critical sectors and environments
✓ Improvements in the way in which 'explainability'[xxii] is approached and improved through practical testing and repeatable techniques
✓ Discovery and learning from what works well and not so well; including limitations
✓ Production of multiple pieces of guidance written by demonstrators as their projects reach completion included in the Body of Knowledge for others to learn from
✓ Asking 'new' or 'better' safety-related questions e.g., that can lead to improved procurement of AI systems from manufacturers/technology providers by organisations
✓ Learning what are the right questions to ask with greater authority
✓ Having to adapt to the realities of the pandemic and pivot to another methodology

A further indicator of success is the significant leverage, £14,048,906[xxiii] that has been achieved by the demonstrators. This encourages positive conditions for enduring change and provides contributory capacity for further research and collaboration in the field of safety assurance.

Testimonials from willing contributors suggest that demonstrators have been a positive experience providing them with the conditions, means and encouragement to develop reports, papers, methodologies of work and their safety assurance expertise. Here is just one example followed by some extracts from case studies completed in 2022.

> *"The demonstration project enabled us to produce a report highlighting regulatory and legal challenges that need to be dealt with to enable the operation of autonomous and remotely controlled ships in UK waters. This report has been well received and been cited by the Maritime and Coastguard Agency and also by various academic works. This report is freely available to all interested parties (*Baris Soyer, Professor of Commercial and Maritime Law, Swansea University, PI on the Swansea University demonstrator)

Exhibit 1: Manufacturing sector demonstrator. Case study acknowledgements: Nicholas Hall, HM Principal Specialist Inspector (Advanced Automation and Cyber Security), Health & Safety Executive and Dr James Law, Director of Innovation and Knowledge Exchange at Sheffield Robotics and leader of the Collaborative Robotics Group

The CSI:Cobot[xxiv] demonstrator project was designed to explore how the safety of cobots could be assured to support increased productivity in manufacturing. Safety and trust issues were hindering their deployment so this project sought to demonstrate how novel safety techniques could be applied to build confidence in the deployment of uncaged cobot systems operating in spaces shared with humans. Details of the research, outcomes, guidance and papers produced are found on the AAIP website[xxv]. Contributors to this project reflected that the experience had impacted them personally and will have wider effects as a result of their work and influence in future too.

Nick Hall from the Health and Safety Executive has been an active participant and critical contributor to the demonstrator project through weekly project meetings where he made suggestions to make assurance 'more regulator friendly' and educated peers on risks and hazards. He recommended putting features into the digital twin to make it risk-based for real-world application. This meant collaborating with c15 people in teams leading on sensing, cyber security, digital twinning and safety synthesis. He was also a judge at a challenge-based workshop where students and recent graduates were able to interact with the digital twin.

In September 2022, he designed and delivered a workshop that brought 20 people together from the HSE (including the head of manufacturing policy), industry (including technical innovation specialists, integrators and safety specialists from large firms) and CSI:Cobot / university researchers.

"The demonstrator project has helped me understand the potential for lower risk applications in complex environments. The demonstrator has shown what could be possible from an assurance perspective and developed a framework that can be compared with other projects to help build consensus."

Nick's learning and experience is likely to be a contributory factor in his thinking and technical work to develop a range of standards with collaborators from around the world.

"As a standards writer you draw on lots of influences, so the AAIP experience may contribute to work I'm involved with around the new Industrial Robots safety standards ISO10218 parts 1 and 2 and the new Autonomous Mobile Robots standard which I'll be working on from next year."

Dr James Law (from the University of Sheffield, the academic lead on CSI:Cobot) agrees that having regulator involvement early on has been beneficial alongside industrial and academic partners. This fosters the conditions for building of trust and confidence through shared, practical experiences. The digital twins helped identify hazardous occurrences through a sophisticated simulation approach.

"By building safety standards into the twin, stakeholders can check how a system meets existing regulation. We have developed a theory of how this could work, and are currently creating a partial implementation."

Looking to the potential enduring effects of this demonstrator James feels that the development of a framework from this collaboration opens up the opportunity to further enhance safety and regulator understanding across a wide range of collaborative robotic processes beyond manufacturing.

More generally, the consultants note that the CSI:Cobot demonstrator project forms just one part of a wider portfolio being managed by the AAIP team as part of its research pillar known as SADA: Safety assurance of decision making in autonomous systems.

At the AAIP collaboration workshop in November 2022, it was reported that there had been multiple achievements aligned to the SADA pillar including fundamental research[xxvi], research grants leveraged beyond the AAIP funded activity[xxvii], work that they feel has influenced standards and research agendas[xxviii]- and contributions to AAIP demonstrator projects exploring shared control in autonomous driving, safe robots for assisted living, and the assuring safety of cobots such as in the example above.

Emerging from this work is learning around Decision Safety Analysis (DSA), a process that addresses the challenges and limitations of traditional safety analysis methods, and also discoveries linked to safety controller synthesis i.e., discrete-event safety controllers for mobile robot - human collaboration/interaction.

The AAIP team realised that new ranges of collaborative robots are being introduced, operating alongside humans without safety cages. There were safety concerns as these robots can be deployed on real industrial processes and operate in close proximity to

untrained and frail users in social care. There was a lack of best practice examples and so demonstrators like CSI:Cobot could help understand how human safety can be ensured and assured such that it inspires confidence in stakeholders.

The team conclude that the CSI:Cobot demonstrator has delivered a multi-stage approach that can be generalised to a broad range of mobile-cobot scenarios, and many of its activities can be automated. It had also successfully tested the safety controller in the digital twin.

Exhibit 2: Automotive sector demonstrator. Case study acknowledgement: Lars Kunze, Departmental Lecturer in Robotics in the Oxford Robotics Institute (ORI) where he leads the Cognitive Robotics Group (CRG)

Lars has been a key collaborator in Sense-Assess-Explain (SAX)[xxix]: building trust in autonomous vehicles in challenging real-world driving scenarios demonstrator project which has explored how autonomous vehicles can be developed that can explain the decisions they take – to the driver, but also regulators, accident investigators and systems developers.

> *"Explainability is critical if we are to gain the trust needed for autonomous vehicles to be used."*

The SAX project used a technique called commentary driving, which Lars explained is *"where you describe what you see, what you anticipate happening then how you reacted to these situations."* This commentary was analysed by the team alongside data being gathered by the vehicle's systems.

Methods for interpreting and representing observations of the environment in human-understandable terms were extended through the work. This has also improved the way that traditional sensors (e.g., cameras) and lasers are used in complex and rare traffic situations. The resulting ten hours of commentary (part of a wider dataset of 140 hours covering over 3,700 miles of on and off-road driving) was a valuable output from the project. It has the potential to be used to validate design and assurance capabilities.

> *"The dataset could be used for validating and showing the system capabilities, for example, localisation across different areas. We wanted to look at a rich variety of environments from the city centre of London to the highlands of Scotland. The dataset including annotations and external sensor data will help to validate and assure performance."*

As a result of the project, Lars was contacted by the European Commission and is now part of the expert group focusing on explainability for automated and autonomous driving for the Commission's Joint Centre for Research.[xxx]

Exhibit 3: Health sector demonstrator. Case study acknowledgement: Dr Nigel Rees, Head of Research and Innovation, Welsh Ambulance Services NHS Trust (WAST)

Nigel was a key collaborator in the 'ASsuring Safe Artificial Intelligence in Ambulance Service 999 Triaging' (ASSIST)[xxxi] demonstrator project which sought to improve the chances of surviving an out-of-hospital cardiac arrest by using AI to support ambulance service call centre staff. The demonstrator project adapted an existing Corti[xxxii] AI platform, which has been piloted in Copenhagen, for use within WAST.

ASSIST comprised three work packages involving different collaborations focusing on safety assurance which involved accessing documents, safety assurance reports, looking at data and coding; ergonomics and interface/interviews with the call takers working in the contact centre to help understand and specify the operating environment for the AI system and determine safety assurance requirements at the clinical system level; and the socio-political stakeholder engagement including the delivery of papers and presentations to various audiences including regulatory and standardisation bodies and from ambulance services nationally.

Nigel explains that the demonstrator's impact will continue for years to come and is:

> *"Much bigger than the adaptation of a tool for early recognition of cardiac arrest. It has given us the keys to unlock the potential for future AI."*

There have been numerous benefits.

The project has brought academics, WAST and CORTI together to share expertise and different disciplines across AI, information governance and data. The learning provides significant contribution to the Body of Knowledge for assurance cases of AI in critical sectors, for example, around defining the operating environment for AI using SEIPS[xxxiii] – a systems approach and its application in an ambulance service context. The collaboration has led to the establishment of a community of practice and the learning being shared is influencing legislation and policy.

Added value

Health and Safety Executive

Nick reflects on his experience of the CSI:Cobot project.

> *"On a personal level, being involved in the work of the AAIP has helped me build links, improve my competence to understand a complex robot system using machine learning that I had not been exposed to before, and appreciate the different interactions involved and what can be ringfenced.*
>
> *Without the AAIP engagement I would probably have arrived at a similar position in my thinking, however, I think the AAIP has accelerated my development along this path, and it has been achieved in spite of competing demands.*

The AAIP has got scale, a range of different demonstrators across sectors and cross-learning is really beneficial to regulators so accessing the AAIP's Body of Knowledge is valuable too."

University of Sheffield

James reflects that the embedding of AAIP's theoretical framework for regulatory information within the system was of additional value because this could lead to automation of safety analysis, which is currently done 'slowly and painstakingly' by hand.

University of Oxford

Lars reflects that for the SAX project COVID brought significant challenges, but that the AAIP provided flexibility and support, which alongside their cross-domain knowledge was vital.

Through his role as a Programme Fellow, Lars also became aware of AMLAS. His current collaboration with York on the Responsible AI for Long-term Trustworthy Autonomous Systems' (RAILS) project includes work to extend this document, taking the guidance beyond implementation. [xxxiv]

"AMLAS is really good, well written and specific. In RAILS we will consider what happens if there are changes in the system once it's been deployed."

Without the wider AAIP infrastructure, the research would have had less impact, which in turn would have meant the collaborations that followed would have been far less likely.

"Without the AAIP, being in the network and the visibility that created would have been missed. The Programme has been great at facilitating and disseminating our papers and enabling opportunities to present at conferences."

Welsh Ambulance Services NHS Trust (WAST)

Nigel reflects on the future impact of ASSIST as well as what has been achieved already.

"We might have procured the system and learnt as we go as issues emerged, which isn't ideal; or we could have discounted it altogether and totally missed the opportunity to explore the technology in our operational context which would have been worse. AAIP provided not just funding, but gave us confidence, networks, insights and expertise across disciplines we might not otherwise have accessed as easily including IT, engineering and human factors."

Where next?

HSE
Nick reports that many parts of the HSE are interested in AI and autonomy as are other regulators who could benefit from the sort of collaboration he has enjoyed with AAIP. HSE are currently developing policy on assurance of autonomous systems that to some extent will have been influenced by his mutually beneficial relationship with AAIP.

> *"I'm open to look at work that would test any industry guidance produced e.g., safety assurance guidelines applied to real equipment including the safety case and let regulators really stress test it."*

Sheffield Robotics
New grant proposals have been inspired by the experience to continue the collaborative robot journey. Interestingly, the results from the CSI:Cobot project are too far forward for industry to adopt at present.

> *"This way of working is very novel and a big change in how people approach safety, so building confidence will take time. CSI is a step too far; industry needs to see more actual examples."*

Instead, to build the confidence required, further, smaller demonstrator projects involving real hardware in actual manufacturing spaces are being planned. In five years, with the AAIP's support, James hopes to be able demonstrate the real life potential of the approaches pioneered here.

> *"Over time, I think the digital twinning tool will provide a means of actually pulling through some of that other research and helping generate financial value too."*

Oxford Robotics Institute
The collaboration with AAIP and the University of York continues through further research. Richard Hawkins[xxxv] and Lars are co-leads on a project for the Autonomous Systems Programme. Another follow-on project with York, RAILS, extends their investigations to maritime and the use of drones. Lars also identifies the opportunity to look more deeply into international standards and regulation.

WAST
Nigel is excited for the future as a result of ASSIST and the wider networks being created.

> *"Our Chief Executive is keen on the potential for deploying these systems and technologies further down the line at operational levels, and it is not unreasonable to suggest that ultimately anyone making a call to our centre – c 600,000 per year – could benefit from the deployment of safe AI. It starts with cardiac arrest detection but can grow from there."*

Other contributors to this report referenced work they were now going to be able to take forward as a consequence of their involvement with the AAIP demonstrator projects too.

"Working with the AAIP team resulted in novel methodologies for the safety assurance of shared control in autonomous driving. It also enabled me to develop my own knowledge of the state-of-the-art in this area. I'm now taking this forward in a new project that aims to develop principled approaches and tools for assuring and demonstrating accountability of safety-critical autonomous systems with respect to laws and regulations." (Lu Feng, Assistant Professor, University of Virginia Co-Investigator on the Safe-SCAD[xxxvi] demonstrator)

"AAIP has provided us enormous opportunities to work with colleagues who share common research interests in robotics, AI, safety and autonomy. A new and exciting research link between UCL and NOC has been established and two RAs have successfully undertaken state-of-the-art research activities with one RA now moving into industry in ocean engineering.

With the support from AAIP, the research team at UCL MechEng can further expand its expertise in marine engineering and further push the boundary of marine autonomy." (Yuanchang Liu, Lecturer, UCL, PI on the ALADDIN[xxxvii] project)

Conclusion

The case studies and testimonials reviewed support the AAIP team's assertion that the *'demonstrator projects contribute evidenced, repeatable techniques for demonstrating the safety of autonomous systems.'*[xxxviii] They have delivered and/or validated guidance that has fed into the Programme's Body of Knowledge.

They have enabled outcomes for participants that have impacted their cognition, attitudes, behaviours and desire to continue collaborating on projects that further the capability to assure the safety of RAS in real-life environments. AAIP has also encouraged the conditions for collaborators to leverage other resources to further research and exploration catalysed or amplified by the demonstrators.

Taken together the demonstrators are making progress towards some of the original outcomes detailed in the Programme's logic model relating to the ability to design for assurance, though there is still some way to go to achieving generalisable results across the domains and jurisdictions.

3.3 AAIP's development of professional education programmes

How have we equipped safety engineers and others with the skills they need now?

The challenge

> *'Working in safety assurance of robotics and autonomous systems requires different knowledge, skills and behaviours, to working with more tradition complex systems. Increasingly, open environments, machine learning and new stakeholders have disrupted the educational model for teaching safety critical engineering.'*[1]

The gap between the level of workforce skills and competencies needed for safety assurance, both now and increasingly in the future, and the pace of technological advance is perhaps the most significant challenge the AAIP can contribute towards. Take up of training has been slower than perhaps anticipated, for reasons not fully understood.

What happened?

To develop and strengthen the skills of those involved with, or regulating these novel technologies (now and in the future), the Programme provides four different types of education:

1. Academic education: a master's level module within the University of York's Critical Systems Engineering MSc[xxxix]
2. Industrial education: onsite or online
3. Research dissemination: conferences, journal papers and workshops
4. Informal dissemination: digital and non-digital outputs

To best create the conditions for success, the Programme's training has to date been aimed at the intersection between autonomous systems, AI / ML, and aligned subjects such as safety, security, legal, ethical, and social.

One of the required outputs from demonstrator projects are case studies and practical guidance aligned to one of the Programme's research strategy pillars. These artefacts are available to other researchers and can also be used in training and education for those developing or working with autonomous technologies.

The difference being made

The insights suggest that:

- Education and training courses and events are affirming and developing knowledge, and building confidence to take informed decisions e.g., to try new things.
- Training courses have targeted a range of staff who were 'close to the use of the technology or product' and who would be able to influence and impact the awareness, knowledge and practice of others.

[1] The sources for this section and 'what happened?' are 'Assuring Autonomy International Programme. A Year in Review' 2019 and 2021.

- Conferences and seminars connect and extend the community and act as a forum to share challenges and formulate research questions.
- Rooting content in the learning and examples generated by the demonstrator projects ensures that learning is research-led, up to date and applicable.

These assertions are exemplified through the following case study extracts.

Exhibit 1 Healthcare. Case study acknowledgement: Sean White, Safety Engineering Manager, NHS Digital

The collaboration with NHS Digital is a mature example of how the Programme works within a sector to upskill the workforce. In addition to three demonstrator projects[2], bespoke training has been designed and developed to complement the national portfolio of training that NHS Digital team has delivered since 2008.

The training delivered has been carefully targeted to increase the likelihood of influencing more widespread adoption. Following strong feedback from a one-day pilot course, more in-depth training courses were developed which targeted small groups of staff who were 'close to the use of the technology or product' and who would be able to influence and impact the awareness, knowledge and practice of others.

> *"We ran this course twice and feedback from learners was that they had a stronger understanding of the state of the art; it corrected misperceptions about AI; and it encouraged them to think differently about how to assure its safe deployment." Sean White NHS Digital*

Case study interviewees uniformly highlighted the expertise that the AAIP team brings, and a refreshing continuity of support too.

> *"The AAIP gave us access to experts – including Ibrahim and Richard whose infectious enthusiasm, and championing of healthcare and safety, encourages success."*

The course was built around the AMLAS guidance, which one NHS Digital delegate noted in their feedback *"Good session, the AMLAS resource is excellent*."[3]

On a different scale were three 'AI in the NHS' national conferences. Held in 2019, 2020 and 2022 they brought together technical manufacturers, human factors specialists, health organisations and regulators all looking from different perspectives at how to use AI safely – discussing regulation and safety strategy, robust assurance methodologies, clinical perceptions and trust in AI, and what can be learnt from other industries.

The AAIP's convening power was also highlighted by a representative of the Health and Safety Executive. Attending helped participants come up with detailed answers to issues that they had clearly been thinking a lot about.

---

[2] 'Safety Assurance Framework for Machine Learning in the Healthcare Domain' (SAFR, 'Safety of the AI clinician' and Safety Assurance of Autonomous Intravenous Medication Management Systems (SAM)

[3] Source: Pilot post course evaluation form 2022 n=5)

> "It gave rise to discussions about what might need to change in standards/ legislation and helped generate ideas for further research." Nicholas Hall, HM Principal Specialist Inspector (Advanced Automation and Cyber Security)

In the second case study below, we consider the short-term outcomes from the MSc module for a safety engineer working in the defence sector.

Exhibit 2 Outcomes from the MSc module

The MSC module is intended to affirm and extend knowledge so that learners can accurately identify, describe and discuss the context, core elements and potential impact of RAS.[xl]

Romas Puisa is a Product Safety Engineer at Thales, working in the defence sector on autonomous vehicles operating on land, sea and air. He has 15 years' experience, and his current role includes the development of conventional safety cases, safety analysis and also research and development projects.

Thales encourage their staff to keep up to date, and as "*York dominates the research in this area",* Romas and colleagues took the module in May 2022.

Romas felt that the course was well organised, and the blend of approaches used was conducive to learning. He felt that there was no particular aspect of the module that was more useful than others, since it was the module as a whole that was considered more valuable.

As well as education and training, Thales is also informed by the knowledge created by programmes like the AAIP.

> *"York's work is helpful here, as every piece of insight helps. The University of York is a part of a wider research community, and contributes to a global, common knowledge that we use."*

Added value

Thales

Romas and colleagues are waiting for a suitable programme to test the AMLAS guidance wholly. However, there are still benefits in the short term. One of the outcomes from taking the MSc module was gaining the confidence to ask more questions and take informed decisions.

> *"Much of education is about confidence. Even though you may not know everything, if you have confidence, you go ahead and you try things and maybe you'll find the solution. Without that confidence, you may not get to those solutions."*
>
> *Taking the module has made our lives easier now, as we know what people are talking about in terms of the safety of autonomous systems. It allows us to make more informed decisions."*

NHS Digital

People make systems safe – but as technology advances, human interaction with these systems 'is more one of guidance than direct command'.[4] The collaboration with NHS Digital has created training that includes human factor considerations. This is a departure from previous training developed by NHS Digital.

> *"The CPD we've developed together has added value to our national portfolio, broadened the scope and depth of our training content, ensured a holistic approach (not purely the technology perspective), incorporated human factors considerations and furthered us towards a position where more people can deploy products safely that meet healthcare needs."*

Without the expertise and knowledge of the Programme, the process of gaining trust and confidence in the workforce around the use of AI would be slower.

Marine Coastguard Agency (MCA)

At the November 2022 collaboration event, the team shared an example of how a demonstrator project on remote controlled and autonomous shipping[xli] in part, led to the delivery of CPD training for the MCA.[xlii]

In the absence of the Programme, the course materials would contain fewer examples and learnings captured 'in the wild' in different environments through the demonstrator projects.

Where next?

NHS Digital

The Programme is considering how best to approach education and training for the next phase of AAIP. The current training offer will likely form part of a wider, blended package of support.

Within the NHS, plans are being developed for an introduction to AI e-learning package which could be made accessible to all those working in the NHS. This would supplement the existing NHS Digital e-learning addressing principles of clinical risk management that has been recently launched.

> *"I would like to see understanding of AI grow and skills develop in effective assurance, the AAIP gives us the resources and knowledge to achieve this. The opportunities are immense for AI in healthcare. It could help with early screening and triaging, bedside monitoring and care, indeed any labour-intensive tasks with the benefit of staff being released to conduct higher skilled work.*

[4] Source: 'Assuring Autonomy International Programme. A Year in Review' 2018

AI is never going to replace nurses and clinicians, rather supplement or compliment what they're doing to deliver care to patients, but we need to build trust in AI and there's still a long way to go."

Influencing workforce training strategies is an important example, with healthcare being perhaps the most mature. As we have read in chapter 1, supplementary AMLAS guidance will be on the NHS Digital website from 2023, and potentially accessed by thousands of professionals and a still wider reach through the Integrated Care Boards (ICBs) in time.

The training and education delivered by the Programme is considered to be technical and detailed. Others, for example another of the AAIP's collaborators, Trusted Autonomous Systems (TAS) in Australia[5] has taken another tact and developed introductory level information to raise the baseline level of awareness amongst stakeholders including regulators and industry. TAS's planned focus on regulation and regulators make them natural allies and future collaborators for AAIP in order to effectively engage this critical set of stakeholders.

In another context, manufacturing, an approach pioneered through the two phases of the CSI:Cobot project was the use of 'digital twins'. This approach has a number of potential benefits, one of which is linked to training. The digital twin enables people to (virtually) see, try out and test technology. Incorporating VR into training or as a part of a suite of e-learning products could feature in the next phase of AAIP.

"Being able to interact in the physical world with a virtual safety representation is very useful in building confidence." Dr James Law, University of Sheffield

Learning points

The initial phases of gaining traction and influence requires education and training recipients to be skilfully selected.

The preferred face to face training method is considered to be optimal by the team, but does limit the numbers that can be reached. It has also proven a challenge to recruit training roles into the team. Take up of training, in general, has been slower than expected, and may be due to the lower maturity in AI and safety assurance in the sectors chosen.

Capturing the 'what next?', from events, conferences and the downloading of guidance materials will provide further evidence of the reach and take up of the Programme's safety assurance methodologies. Evaluating the influence of guidance on workforce training strategies within NHS Digital will provide valuable insights into bridging this key gap more widely.

The inclusion of more real-life examples and challenges (as opposed to content based on foundational research conducted earlier in the AAIP programme) within the MSc module would make it more relatable to participants seeking to apply the learning in their roles.

[5] https://tasdcrc.com.au/

The materials developed to date will require updating to include more examples and learnings from more recent and future demonstrator projects.

Conclusions

The education and training developed by the Programme, informed and shaped by the research and real-life learning and guidance generated by the demonstrator projects, provides learners with both the methodologies and examples necessary to more confidently work on safety assurance and using the guidance, a common language to share and bring others along.

Case study evidence and limited survey feedback suggests that the specific skills needs of safety engineers are being met, but further evaluation will be required to substantiate this. How to bridge the gap between the growing number of stakeholders who will benefit from training, and the team's capacity to deliver this to their required standards, will be a key challenge for AAIP 2, since the current targeted approach can only achieve scale and reach indirectly.

3.4: AAIP's engagement of the industrial, regulatory and academic communities

Challenge

While there were networks for AI or robotics developers, and well-established groups of safety specialists, there was no international community specifically for the assurance of safety in RAS prior to the AAIP.

Consequently, without a system or structures for collaboration, the pursuit of common ambitions around safety assurance could not be guaranteed. There was less likelihood of finding a common language (such as provided by AMLAS for example). And with that, the ability to achieve consensus or acceptance at the pace required to keep up with RAS technological developments was also more limited.

What's changed

In order to address the challenge, the AAIP has developed exciting collaborations and works with a wide range of stakeholders, just some of which are detailed here:

- BSI – standards for AVs
- Energy – safety of AS on solar farms
- HSE – safety of factory automation
- MCA– continuing professional development
- Mining – UAVs for mining operations
- NHS Digital – adaptation/adoption of AMLAS

• Rail – safety of perception in urban transport

• Start-ups – safety of adaptive control algorithms.

The Programme is also part of a set of networks, task and strategy groups.[xliii]

The difference being made

The AAIP Programme has made some difference with regulators, while acknowledging that *"regulations change slowly"* (November collaboration workshop). The potential reach of regulators through standards and regulation is greater than individual businesses.

> *"I think we have made more progress in the assurance space, but there's more to do in the regulatory space. One of the challenges for regulators is the technology, so getting them to interact intelligently with manufacturers is a challenge. We are hosting a workshop with regulators to help shape standards in the regulatory framework." (AAIP team, collaboration workshop, November 2022)*

For example:

- ✓ The BSI Base Document 'Safety Assurance of machine learning for automated vehicles'
- ✓ Dstl 'Biscuit Book', titled 'Assurance of AI and Autonomous Systems'
- ✓ In healthcare, BS 30440 Validation framework for the use of AI within healthcare – Specification' and a Review of the AMLAS Methodology for Application in Healthcare'

The insights suggest that:

- ✓ The development of awareness, interest and trust in each of these communities
- ✓ Skilful engagement, creation and nurturing of multi-disciplinary teams is a key enabling factor to address the complex challenges to the safe introduction and adoption of RAS
- ✓ How the Programme's cross sector perspective and approach can "disrupt" and enhance safety structures
- ✓ The nature and extent of collaboration that sets the Programme apart. The depth of collaboration creates a culture of openness, and trust where questions can be raised.
- ✓ A shared experience of being part of a demonstrator project team leads to both further collaboration and the creation of influence.
- ✓ The AAIP has provided the reasons and the funding to bring together parts of the international community
- ✓ A shared voice that connects the Programme outputs to a potentially wider audience
- ✓ The Programme has prompted participants to develop new ways of thinking, and for some, a shift in their attitudes towards safety of RAS
- ✓ Participants appear invested and emotionally connected to this work.

These assertions are exemplified through the following case study extracts and testimonials from the international AAIP community.

Exhibit 1: The extent of collaboration. Healthcare. Case study acknowledgement Professor Tom Lawton, Bradford Teaching Hospitals NHS Foundation Trust

Professor Tom Lawton is a critical care physician and consultant anaesthetist, and Head of Clinical Artificial Intelligence at Bradford. His collaboration with the Programme includes a project exploring how artificial intelligence can help predict the optimal time for ventilator extubation[xliv].

Tom's role to help bridge academia and industry with real-world clinical perspectives means he can benefit from, and contribute to, the AAIP as a Fellow and research collaborator. Over time, the AAIP has helped Tom widen his network of expertise and access to people not just across the healthcare and academic sectors, but also across multiple disciplines.

> *"Collaboration is essential. That doesn't mean we have to work in the same way though, in fact, a vital role of the AAIP has been to create a space where people with different opinions have been able to gather together and discuss their approaches. This mitigates against silos and group think which is where the mistakes happen.*
>
> *The main benefit of the AAIP is access to a community of experts in multiple disciplines so you can explore legal, ethical, technical questions so that safety is 'live' for all, not just something that is produced at the start of a project and put in a ring binder to pull down when you think you need it."*

Tom describes how the collaboration is distinct from others he has experienced, and how it can lead to improvements - 'course corrections' – early enough in the approach.

> *"It's the way that the AAIP promotes regular contact and dialogue rather than allowing different experts to drift away, that makes it different. This is my wider observation – people think they are doing multi-disciplinary collaboration but I'm not sure it's really happening routinely. The AAIP is really good at encouraging strong multi-disciplinary approaches meaning you get all those course corrections during a process before it's too late and more expensive to put right."*

Exhibit 2: Engaging and influencing regulators through as a result of a shared experience. Case study credit. Nicholas Hall, HM Principal Specialist Inspector (Advanced Automation and Cyber Security), Health and Safety Executive

The Health and Safety Executive regulates safety across a range of safety critical sectors.

A shared experience of being part of a demonstrator project team has led to both further collaboration and the creation of influence. The choice and calibre of these project teams has been a key enabling factor.

> *"I always thought a shift in safety philosophy would be needed to accommodate autonomous equipment, because new methods will require new control measures to maintain safety."*

More generally Nick reports that many parts of the HSE are interested in AI and autonomy, as are other regulators, who could benefit from the sort of collaboration he has enjoyed

with AAIP. He thinks attitudes are shifting from 'AI + safety = no thanks' to a growing understanding that there is a need to develop along this route. Assurance is critical to that path. Consequently, HSE are currently developing policy on assurance of autonomous systems that to some extent will have been influenced by his mutually beneficial relationship with AAIP.

Exhibit 3: Providing the structures so the Programme's shared voice is more powerful. Automotive. Case study acknowledgement: Professor Dr Simon Burton, Research Division Director at Fraunhofer IKS

Simon is a longstanding, valued collaborator with the Programme, who has in turn connected others to AAIP. Key to the AAIP is its ability to connect subject matter experts across a wide range of sectors and disciplines.

> *"The Programme brought us together. It is such a huge topic, with so many actors involved and so much literature being published. With a wider net to throw we can get a good view of what's going on. As a Programme Fellow, I can learn from what's going on in UK industries."*

The Programme Fellows scheme also adopts a different approach by inverting the usual way in which industry accesses the expertise within universities.

> *The other thing that worked really well was the concept of the Programme Fellows, because it allowed people to come from industry and feel somehow intimately involved as part of the team. Basically, they've turned the usual model on its head. For me, I'd normally be paying the university to come and do work for me. It was a completely different type of approach and it's been successful."*

Crucially, the Programme provides a bigger platform for sharing the artefacts of these shared endeavours; the foundational research, the varied outputs from demonstrator projects and crucially, the detailed 'how to' guidance that could be seen to represent the sum of all this investigation and learning, the keys for others to use.

Between 2018 and 2022, the programme has produced over 170 papers which have led to more than 2,000 citations.[6] The Programme has led to increased recognition for the protagonists, and created the conditions for further collaboration.

In another example from the automotive sector, another collaborator, Dr Lars Kunze, Departmental Lecturer in Robotics at the University of Oxford, reflects on how his involvement with AAIP has helped his career and connected him to other sectors.

> *"This has brought recognition to my career. And it was an opportunity to link my research around space exploration to validating the systems used on Mars rovers."*

[6] Source: Programme monitoring data reviewed in October 2022.'

Exhibit 4: International engagement. Case study acknowledgement: Rachel Horne Assurance of Autonomy Lead / Director of Autonomy Accreditation - Maritime at Trusted Autonomous Systems.

Trusted Autonomous Systems (TAS) is Australia's first Defence Cooperative Research Centre and delivers research into world-leading autonomous and robotic technologies to enable trusted and effective cooperation between humans and machines.

TAS recognises that autonomous systems cannot be operationalised without appropriate ethical, legal and regulatory infrastructure being in place. Rachel Horne is the Assurance of Autonomy Activity Lead, and is delivering initiatives under these activities.

The assurance focus aligns well with the AAIP– which also highlights the importance of having effective regulation that can keep people safe, keep pace with the technology and not over-regulate to stifle innovation.

> *"The AAIP team at York are doing world leading research and are at the forefront of the field that they're working in, and that we're also seeking to work in. We don't have any projects specifically looking at how to actually do assurance. For example, what should those assurance methodologies look like? And certainly not to the level of detail that the AAIP have produced. For us, it's a no brainer to collaborate with such an organisation, especially one that's also funded for the public good."*[7]

The team at York delivered a series of three public webinars to TAS's stakeholders, for example one on the AMLAS guidance and another concerning the ethics of assuring AI.

> *"The idea was to draw on the expertise of the AAIP team, and especially John McDermid, and find topics that were of interest to our TAS stakeholders. It was simultaneously more exposure for the AAIP team but also a really a good way of giving our stakeholders access to world leading expertise."*

Added value

Fraunhofer IKS

Without the AAIP community, research outputs would struggle to gain the same traction.

> *"I see a very broad, interdisciplinary consensus that can have more of an impact and achieve greater levels of acceptance."*

More fundamentally, without multidisciplinary teams, the complex challenges posed by assuring the safety of autonomous vehicles would potentially not be addressed as effectively.

[7] Rachel has also written a blog about the importance of collaboration - https://tasdcrc.com.au/collaborating-on-regulatory-environment-for-novel-autonomous-vessels/

"We worked with a philosopher, Zoe Porter, to co-write a paper that set out the ethical dilemmas in a qualified, expert manner. It created a bridge between the technical world and the moral and ethical one. Engineers cannot be blind to these topics, but they can't solve the ethical dilemmas. We hit a glass ceiling, from an engineering perspective, trying to answer the question 'how safe is safe enough?'", unless we take a wider interdisciplinary perspective."

Without the cross-disciplinary approach, and how it has been implemented and refined, the resulting research and guidance from the work would arguably be less convincing.

"We wouldn't have the novel, interdisciplinary ways of thinking. Instead, we would all be working independently. My work would be seen as a lone voice."

Having an international community to refer to and involve has enabled the Programme to peer review draft guidance prior to launch. As the comment below illustrates, the international community extends the reach of the AAIP, broadens the scope of potential collaboration and provides a larger audience that have an interest in the outputs being produced and made publicly available.

The structure employed by the Programme provides the means to not only test and explore, but also to develop and (for the future) to scale up.

"The Programme allows ideas to be tested and workshopped. We determine if the idea is viable, but crucially, then we have the critical mass to actually do something with it.

I see a very broad, interdisciplinary consensus that can have more of an impact and achieve greater levels of acceptance." (Professor Dr Simon Burton, Fraunhofer IKS)

University of Liverpool

New ways of thinking were also highlighted by another member of the AAIP international community.

"Working with experts with diverse skills and experiences has led me to new ways of thinking. The research we've worked on together has benefited hugely from this diversity. I'm using this knowledge to co-design and co-evaluate functional, social, legal, ethical, empathetic and cultural properties of AI systems." Dr Xingyu Zhao, Lecturer in AI, University of Liverpool, AAIP Fellow

KTH Royal Institute of Technology

Being able to convince others to join and fund follow up research was also made by another AAIP collaborator in relation to developing more compelling cross-disciplinary proposals.

> *"AAIP put me in contact with other researchers focused on autonomy, which had complementary perspectives to mine (legal, policy, etc.) that I both enjoyed hearing about, and which gave me a better understanding of various stakeholders I meet in my own work. I think this has helped during my work on other research proposals, making it easier to integrate other organizations in research "narratives".* Dr Fredrik Asplund, Assistant Professor, KTH Royal Institute of Technology
> PI on the BOAUT project

Health and Safety Executive

The choice of high calibre experts from different fields amplifies the influence created and better creates the conditions where learning from one sector can inform developments in another.

> *"The AAIP has got scale, a range of different demonstrators across sectors and cross-learning is really beneficial to regulators so accessing the AAIP's body of knowledge is valuable too." Nicholas Hall, HSE*

The collaboration is relational and reciprocal

> *"Having John McDermid OBE joining the HSE as a non-Exec Board member has also been helpful to raise the profile of AI. Challenges are time and making your voice heard within the HSE where all 2,500+ employees are engaged in valuable work."*

University of Sheffield

Without the Programme, put simply opportunities would have been missed. With technological development, timing is critical. Dr James Law (University of Sheffield) says that without the engagement of the AAIP, "we'd have missed out on much of what has been achieved".

Which includes a plan for a spin out venture to maximise the potential of the digital twin approach pioneered in the CSI: Cobot demonstrator project.

University of Oxford

> *"Without the AAIP, being in the network and the visibility that created would have been missed. The Programme has been great at facilitating and disseminating our papers, and enabling opportunities to present at conferences." (Dr Lars Kunze, University of Oxford)*

Trusted Autonomous Systems DCRC

> *"The biggest thing that the collaboration has done, is it showed us how someone else has tackled the same broad issues that we're tackling. The AMLAS framework, for example, shows us what you could achieve with the time and the funding and the effort to do that technical task. And we haven't tried to do that ourselves in that way – but without the AAIP we might have.*
>
> *Perhaps our next project might build on AMLAS, leapfrog rather than reinvent." (Rachel Horne)*

Where next?

Bradford NHS Trust

Professor Tom Lawton highlights that there is much more work to be done to realise the promise of robotics and autonomous systems.

> "The use of AI to help inform decisions when a person can come off a ventilator is a project close to my heart. Many attempts have been made to work out rules of thumb and algorithms to predict the optimal time for weaning. However, so far nothing has been better than the human eyeball so I'm motivated to continue that work and see how AI can try to make these predictions safely and optimally including the counterfactual explanations. Approved funding from the AAIP enables us to continue this exploration.
>
> Secondly, in a new project jointly funded by the University of York and the Medical Protection Society Foundation we will be looking at how AI and humans can work together using novel simulation techniques which is exciting for the future too."

Fraunhofer IKS

For the team at Fraunhofer IKS, the collaboration with the Programme continues. In 2021, Fraunhofer funded a new, related 3-year project with the University of York on autonomy and safety. The continuity of collaboration being a key enabling factor.

> "It's really about building on these relationships. This project fits nicely into the programme and builds on and is enriched by those collaborations. As the relationships were already developed, we hit the ground running instead of having to do the usual lengthy setup, building up to an intense last six months."

By helping regulators understand this fast-moving landscape, the Programme has already reduced uncertainty. One of Simon's suggestions for the next phase is to use an approach called 'regulatory sandboxes' where assurance specialists are part of multi-disciplinary teams working in a tightly defined area of focus. Nicholas Hall from the HSE makes a related point:

> "I'm open to look at work that would test any industry guidance produced e.g., safety assurance guidelines applied to real equipment including the safety case and let regulators really stress test it."

A complementary second suggestion[xlv] is to also move to bigger projects that introduce and test the safety of these new technologies in real situations to transfer the results achieved towards commercially viable systems[xlvi] This will require a further expansion of the collaboration and increased leverage by the Programme to work at scale.

Thirdly, with the challenges now better defined, there remains numerous research lines of inquiry to pursue – subject to suitable investors, collaborators and early adopters being found.

Trusted Autonomous Systems DCRC

In 2023, Rachel Horne is hoping to set up a new TAS business unit, called 'Trusted Autonomous Systems Advisory.' Their offer will be to improve regulatory frameworks, continue to raise awareness and knowledge, and provide direct assistance to operators and other third parties going through the regulatory process.

In considering how the business unit might work, Rachel was seeking opportunities that would dovetail with AAIP and also an opportunity for the international community.

> *"I would love to look at assurance, and actually do something more tangible, how you actually get documents into people's hands would be amazing. There is also scope to show global collaboration, perhaps through a demo event where we all come together to talk about the issues, maybe have a demonstration of some different technology."*

Conclusion

The multidisciplinary collaborations made possible by the AAIP is one of its foremost assets; creating the conditions for complex challenges to be tackled, knowledge and connections extended, and the results to be more credible / accepted, compelling and widely shared and potentially applied than would otherwise be possible.

Through embedding safety thinking into research, further collaboration, at different scales, means the opportunity for enduring change is created which brings forward all the potential that greater use of RAS could bring.

3.5: The wider landscape

How have we changed the landscape in which autonomous systems are developed?

The challenge

The landscape in which autonomous systems are developed pre-AAIP lacked a consistent focus on the assurance of safety. At this time, neither was there an international community that coalesced around the complex challenges of assuring complex systems. There was no single repository, nor curation for the safety assurance of RAS prior to the AAIP.

The University of York was motivated to lead the Programme as it was seen to be a natural extension of more than 30 years' work on safety in complex systems.

There were around 80 bidders to Lloyd's Register Foundation to lead or be part of the Programme, evidence of at least intent to be part of this landscape.[xlvii]

Awarding the entire Programme to York was a surprise to some, but, with hindsight, is considered to have sent a message to the robotics community that a safety focus was critical, and not in place.

Since the start of the Programme in 2018, there has been greater investment in AI. The example below comes from the Engineering and Physical Sciences Research Council (EPSRC).

Since 2016:
The Artificial Intelligence Technologies research portfolio has increased from 1.86% (£56m) of EPSRC's total research portfolio in 2016 to 3.7% (£128m) in 2022.
Similarly, the AI training portfolio has increased from 2.05% (£31m) in 2016 to 6.41% (£102m) in 2022

In the UK, the application of Automation and Robotics was identified as offering a ten0-year Value at Stake of £183.6bn to the UK economy.[xlviii] A 2017 study identified potential productivity improvements of 22% if the UK invested in automation in line with leading nations.[xlix]

What's changed

The insights suggest that:

- ✓ The Programme has influenced funding decisions, leading to increased investment in safety assurance of autonomous systems.
- ✓ AAIP has developed and given the landscape a structure that was previously lacking
- ✓ York's ability and motivation to lead a cross-sector Programme matched the requirements of Lloyd's Register Foundation.
- ✓ Safety expertise was missing from the landscape for robotics and autonomous systems.

These assertions are exemplified through the following case study extracts and market insights provided by the international AAIP community.

> "York wasn't a big player in robotics so for some people it was a real shock that York led this bid and got the whole investment from LRF. LRF decided all the expertise could be assembled by York. This sent a message to whole robotics community that they needed something that they didn't have – which was the safety element." (AAIP team reflection, September 2022).

Exhibit 1: Creating the building blocks for safety assurance of autonomous systems Case study acknowledgement Professor Dr Simon Burton, Research Division Director at Fraunhofer IKS

Simon considers that a key achievement from the AAIP is bringing greater clarity to this landscape.

> "It was not a coincidence my path and York's kept converging. Because of the progress in the area of artificial intelligence, all of a sudden, it looked like the world was opening up to all these new possibilities. At the same time. It was opening up a whole new raft of questions, which traditional safety engineering wasn't really built to answer."

The discipline of assurance of autonomous systems is evolving and beginning to take shape.

> "I call this new field of research 'safety assurance under uncertainty'. Structures are beginning to form whereas at the start of the AAIP it was vague, like everything was shrouded in cloud. We are trying to understand what works, what doesn't work and how to approach these issues. The next phase, I think, is really about trying to feed this into reality."

Exhibit 2: Creating influence that leads to greater investment in safety assurance of autonomous systems. Case study acknowledgement: Dr Helen Niblock, Head of Regional Engagement (NE, Yorks & Humber) Engineering and Physical Sciences Research Council

The Engineering and Physical Sciences Research Council's (EPSRC) relationship with the AAIP team at York is rooted within the UKRI mission "to convene, catalyse and invest in close collaboration with others to build a thriving, inclusive research and innovation system that connects discovery to prosperity and public good."[l]

Dr Helen Niblock is Head of Regional Engagement (NE, Yorks & Humber) at the EPSRC. She works across the region, with research and innovation stakeholders to understand their research and innovation activities, identify synergies and to make connections. Helen was previously at the University of York.

Funding for research and application projects using artificial intelligence has increased significantly over the lifetime of the AAIP Programme. York's influence was particularly apparent in the Trustworthy Autonomous Systems Programme which launched in 2020.[li]

The Trustworthy Autonomous Systems Programme (TAS) is an EPSRC led UKRI programme of £33m. TAS' vision for this programme was to enable the development of socially

beneficial autonomous systems that are both trustworthy in principle and trusted in practice by the public, government, and industry.

> "York influenced the topics of the nodes within the programme leading to an open competition as to who got funding. [Members of the AAIP team] were involved in conversations with EPSRC during the development of the TAS business case and this involved the topics of the nodes. These conversations influenced the strategy and priorities of the programme."

Following an open competition, York was one of the recipients of this £33m programme. Specifically the £3m Autonomous Systems Node in Resilience, led by Professor Radu Calinescu. York are also partners in the Verifiability node.[lii]

Added value

Fraunhofer IKS

Professor Dr Simon Burton asserts that the Programme approach and rigour has led to increased clarity and definition of the safety assurance landscape.

> *"If you use qualified language, in other words, language based on a set of formal definitions, we can begin to actually understand what complexity actually means and how that relates to the types of safety challenges we have. It means we can also create new perspectives that help structure our approach to achieving safe systems. I'm applying that approach in my work here."*

Simon sees the Programme's wider, cross-sector perspective as becoming more important because of its ability to disrupt:

> *"There are different paradigms for safety in different sectors. Each has their own paradigm, and they are reaching the limits of those. Structures have become calcified in these industries – particularly in the language and setting of industry standards.*
>
> *The Programme can then disrupt those staid structures. It's something that the AAIP is really well placed to do because it's got this broad view and looks at the concept of autonomy itself, rather than one particular industry, and the problems associated with autonomy rather than road regulations for example."*

EPSRC

Without the relentless focus on safety assurance from the AAIP, it is unlikely that comparable frameworks to AMLAS and SACE would have been created.[liii] None has the focus of the York team.

> *"There was no framework in place and if York hadn't been doing that work, I don't know who would have done it."*

The Programme's contribution to the evidence base, brought together in the 'Body of Knowledge' and in the guidance documents created, would have proved difficult to emulate

by other means. While speculative, other bidders were likely to have taken a narrower focus.

> *"York's work on safety and policy have made a massive contribution to the wider landscape. The Body of Knowledge has been a big piece of work which has been very useful to academics and professionals as well."*

University of Sheffield

The Programme has also created unforeseen outcomes.

Part of the CSI:Cobot project that is considered particularly exciting, and commercial, is the use of 'digital twins' which carry the potential to automate safety case generation. In addition to providing a mechanism for integrating and testing the team's safety and security approaches, with the regulator the team developed an approach to help identify hazardous occurrences using a digital twin.

The subsequent emphasis on digital twins was actually the fortuitous, unplanned result of COVID. As restrictions prevented access to both laboratories and manufacturing sites, rather than the modest amount of modelling anticipated, the team quickly realised that to move forward, their digital tools needed to be both more ambitious and sophisticated.

> *"We realised that rather than developing a model of a specific process, we were going to need models that could reflect multiple types of processes, robots and systems. Having a digital twinning system was going to be the most useful way to collaborate and test our methods offline and reduce the amount of time required for integration later on. That led us to develop, and continue to develop, a sophisticated digital twinning framework."*

Without the Programme, the scope for cross-sector application would have been diminished.

> *"The development of a framework opens up the opportunity to further enhance safety and regulator understanding across a wide range of collaborative robotic processes beyond manufacturing."*

The expertise of the AAIP team would not have grown, nor added value in the same way without the Programme. York's credentials set them apart.

Where next?

EPSRC

There is an opportunity for the AAIP team at York to reconnect to new senior managers within the EPSRC

More generally, the Programme should reflect on how their work aligns and complements the work of others, for example the Office for AI. AI was the subject of a national strategy published in 2021.[liv] The EPSRC planned investments in AI make it a potential future funder for demonstrator projects from the AAIP. Dr Helen Niblock advised that the AAIP continue

to consider how their work connects with the significant investments planned, including £80million to establish up to eight hubs across foundational AI, AI for Real Data and a number of application areas and new AI Centres for Doctoral Training, in line with the government's announced intention to fund an additional £117m.[lv]

Importantly, the majority of investment will be directed towards how technologies are adopted and implemented.

> *"It's looking at the adoption and translational aspects in those strategies, contributing to its economic and societal value with very high adoption of the technology, adoption of AI that works for everybody and trust is certainly part of that.*

University of Sheffield

The project opens up the opportunity to further enhance safety and regulator understanding across a wide range of collaborative robotic processes beyond manufacturing.

> *"York's work to help industrial organisations navigate the process of building a safety case for robotic system is very interesting. So far, we've largely been looking at the safety of industrial robots because that's where the earliest adopters are for collaborative technology. However, collaborative robots are something you're going to see across all sectors, particularly health and social care, and in areas like logistics, public services, and entertainment as well."*

York and Sheffield's collaboration continues. The final output from the CSI:Cobot project was a joint workshop (September 2022) with the HSE on the safety of mobile robots. At the final project meeting, a new industrial partner stepped forward and requested to get involved. This interest has inspired the teams to develop new grant proposals to continue the collaborative robot journey.

Interestingly, the results from the CSI: Cobot project are too far forward for industry to adopt at present.

> *"This way of working is very novel and a big change in how people approach safety, so building confidence will take time."*

To do this, further, smaller demonstrator projects involving real hardware in actual manufacturing spaces are being planned.

In five years, with the AAIP's support, James hopes to be able demonstrate the real life potential of the approaches pioneered here.

> *"If we can commercialise the basic digital twinning framework, we can then look at working with other partners in the research domain, to bring in new services and tools, like some of the safety analysis tools that were developed in this project. Over time, I expect the digital twinning framework will provide a means of pulling through some of that other research and helping generate real-world impact too."*

More generally, the global market potential for the collaborative robot market is expected to reach $5.6bn by 2027, accounting for 30% of the total robot market.[lvi] Sales of mobile cobots are expected to reach $209m by 2023, r 9.7% of the total cobot market.[lvii] In five years, with the AAIP's support, James hopes to be able demonstrate the real life potential of the approaches pioneered here.

Conclusion

The AAIP is one of a number of actors with a stake in assuring the safety of robots and autonomous systems. Unlike others, their consistent focus on safety assurance is unique, and the team is regarded as being world leaders in this field.

The methodology created by the Programme - termed by one collaborator 'safety assurance under uncertainty', **has the potential to create an enduring legacy.**

The Programme has skilfully influenced decisions on the dispersal of funding in favour of safety assurance.

4 Learning lessons

Chapter 1: Enhance design and assurance practices

How are industrial practices safer because of our work?

Feedback loops with collaborators during the development of guidance has been beneficial to the extent that AMLAS has provided an approach and aesthetic emulated by subsequent published guidance (SACE), and is more widely replicable in future.

In the absence of AAIP some stakeholders feel that they might be able to access other design guidance; have communication with regulatory bodies; and undertake extensive verification and validation, however, they do not feel they could make the same progress, access the expertise that AAIP offers 'in one place' nor at the same pace.

There are some challenges that might prevent targeted audiences for the AMLAS guidance to use it to its full potential including access to training and / or the opportunity to be involved in funded programmes employing ML and/or AI technologies.

The perception that implementing AMLAS is expensive is worthy of further investigation to assess the actual costs which can be weighed against current approaches and consider potential savings; for example, by being able to implement more quickly and effectively. Meantime, the AAIP is proactively investigating ways to develop a proportionate approach to safety case maintenance / revision post-deployment of an AS with cost-benefit in mind.

As the AAIP will only be an indirect contributor to safety practices and processes, i.e. those who have downloaded the guidance bearing the responsibility of using it appropriately and contextualising it for their context, encouraging the sharing of examples of systems that have changed as a result will be useful evidence for the Programme. An AMLAS community

of practice, built around the emerging case studies where guidance is being applied to RAS safety cases, for example would be one forum where users can share, celebrate and learning from one another.

New and innovative safety assurance exploration is being led by the AAIP that will provide valuable additions to design and assurance capabilities in areas that are breaking new ground and generating curiosity amongst the ML community e.g., around ethics and responsibility. These in turn, are helpfully addressing the 'delta' where the distinct properties of AS mean that a different or better approach is required compared to the safety assurance of traditional systems.

Chapter 2 – Validate [the design and assurance capabilities] through translational demonstrators

How have we impacted safety-critical sectors?

Early involvement of regulator(s) in demonstrator projects adds significant value to the experience, learning and outcomes achieved, saving time and money whilst also shaping the different perspectives towards a greater likelihood of consensus and RAS safety acceptance.

The AAIP also learned that there is benefit in having more of their team members embedded in the demonstrators either from the start or earlier on in the lifecycle of those projects. These approaches save time and identify 'the right questions' to ask across the collaborations much sooner than in their absence. It also increases scope to bring in learning from other sectors or demonstrators.

Demonstrator projects did not always meet their intended objectives but have persuaded those involved of the potential of AI in their sectors. Critically, this is not regarded as failure, but a springboard to further collaboration. They have created sufficient time to allow people to discover and / or change their perspectives.

The golden thread of being able to work closely with experts across disciplines is seen as a key success factor in pursuing an ambition to improve design and assurance capabilities, including for example, the inclusion of ethicists in any multi-disciplinary team developing safety assurance of RAS.

> *"It is important to have ethicists in a multi-disciplinary team to make considered judgements about and then ensure the identification of equitable distribution to support top level argument that it is ethically acceptable for the deployment of a system. At AAIP we really do take multidisciplinary working seriously – we don't just talk the talk." AAIP collaboration workshop participant, November 2022.*

The AAIP team feel they learned useful lessons about how to select and invest in demonstrator projects with more confidence over time, noting that sometimes there were unintended consequences (benefiting the investee perhaps more than the AAIP), but in all instances there was a reciprocity in terms of exchanging learning. Investments in demonstrators might not be as large in future based on the learning gained.

Case study participants shared interesting learning too. Nicholas Hall from HSE reminds us that there are other drivers and actors beyond the AAIP that can support successful conditions in this field despite competing demands for time and attention across safety critical occupations and sectors:

> *"It has been helpful to have central government driving the need for more attention and focus, for example creating the Office for AI and publishing the UK National AI Strategy."*

Nigel Rees reflects on the ASSIST Project:

> *"We think that we have learned a lot about 'complexity' and why it was not possible to get the data to train the model as originally intended. It exposed external factors that need to be in the right place to enable these new systems to thrive. We learned that it is extremely challenging to deliver something like this during the biggest emergency we've ever faced, on top of what is always and already a high pressure (emergency services) environment where callers are hearing harrowing conversations in their ears all day and have to make time critical decisions.*
>
> *We think that AI does have potential for our context. Ultimately, we think AI could impact survival from cardiac arrests, and if we can demonstrate that further in future, that has applicability to a far wider set of deployments in the service."*

At an AAIP-hosted collaboration workshop in November 2022, collaborators were working through a process to define a potential set of projects to take forward from 2023 – around two to three projects across four selected domains. This process may usefully test the divergent attitudes observed in this review about the need for large- and/or small-scale real-world projects (subject to external funding) to build on the foundational work achieved by the demonstrators since 2018.

Chapter 3 Develop professional education programmes

How have we equipped safety engineers and others with the skills they need now?

Learning Points

The Programme has developed the first MSc module dedicated to the safety assurance of autonomous systems and machine learning, 'Advanced Topics in Safety' which helps to strengthen the team's authority in the safety assurance domain. Viewed in isolation, the initial level of take up is low, but, with an eye to the future, and as part of a wider, blended set of learning and CPD, the module can make an important contribution.

To understand how the workforce can be trained at scale, there has also been valuable learning in how the AAIP has worked with a large organisation such as NHS Digital to develop customised professional education and training aligned to their wider organisational goals around AI/ML skills and competence. How the Programme established traction and influence through the skilful selection of education and training recipients will likely be replicable in other contexts.

The AAIP team feels that any approach to professional education and training for RAS safety assurance needs to be experiential. This approach, if considered in isolation from complementary approaches to educate the market presents obvious challenges to capacity given the finite tutorial capacity of the AAIP core team when compared to the size of the workforce that may require safety assurance skills now and in future. Since 2018, the preferred training method of the AAIP team is face to face. This is considered to be optimal by the team, but it does limit the numbers of people that can be reached. It has also proven a challenge to recruit training roles into the team. Take up of training, in general, has been slower than expected, and may be due to the lower maturity in AI and safety assurance in the sectors chosen. This set of challenges is addressed in one of the key recommendations made by the consultants as a result of this review (please see recommendations section).

Capturing the 'what next?', from events, conferences and the downloading of guidance materials will provide further evidence of the reach and take up of the Programme's safety assurance methodologies. Tracking the influence of guidance on workforce training strategies within NHS Digital will provide valuable insights into bridging this key gap more widely.

The inclusion of more real-life examples and problems to work through (as opposed to content based on foundational research conducted earlier in the AAIP programme) within the MSc module would make it more relatable to participants seeking to apply the learning in their roles.

The materials developed to date will require updating to include more examples and learnings from more recent and future demonstrator projects.

Chapter 4 - Engage the industrial, regulatory and academic communities

How have we guided the safe development of autonomous systems across the globe?

The engagement and integration of regulators into demonstrator project teams took time to develop, but their presence became a feature of new projects from 2021 onwards.

The approach has helped to safeguard the University of York's leadership status through a time of great uncertainty, and build a dedicated, £35m Institute for Safe Autonomy facility to take the Programme forward.[lviii]

> *"Finding the time for collaboration is hard. What I can say is the AAIP approach definitely works and it's gaining critical momentum. Despite Brexit, which makes it hugely challenging for institutions like York to access funding, they've employed a model that pools expertise and provides a wider surface for the university. York has been very creative in finding ways to engage with industry." Prof. Dr. Simon Burton, Research Division Director, Fraunhofer IKS*

Finding the right language and models to engage regulators

> *"The skill is in finding the right language to involve them in this interdisciplinary work. Finding a set of models, which you can use to structure that work." Prof. Dr. Simon Burton, Research Division Director, Fraunhofer IKS*

The international AAIP community is seen to have been valuable in:

- Refining the guidance
- Being both collaborators and ambassadors and advocates for this way of working.

However, there is significant room to expand the international community, with key gaps identified in China and sub-Saharan Africa.[lix] A challenge is the time for a member of the York team to travel, or vice versa, in order to imbue an approach that can then be contextualised. This challenge is addressed in the recommendations section of the report.

Chapter 5 - The wider landscape

How have we changed the landscape in which autonomous systems are developed?

The AAIP has created an international community where before it did not exist. While York is seen to have brought greater clarity and structure to the landscape, the true size and scope is yet to be defined.

As part of the next phase of AAIP, the team may wish to undertake a stakeholder mapping exercise to identify those whom the Programme is best placed to influence – against its five research pillars - as well as those that are critical to the AAIP's continued success.

The AAIP approach is seen to work for three reasons:

1. York's reputation as a centre of excellence in this area
2. The quality of the project management
3. How the programme has involved and worked with industry.

> *"York has the pedigree and is seen as having the right level of expertise, so the trust is there that in York there was an organisation to address these issues. The second aspect is the execution. Reputation alone doesn't make a successful project. It's the execution and that's where Dr Ana MacIntosh has done an excellent job. I think she's bought the right mentality and most certainly the right energy, ambition and drive."*
> Professor Dr Simon Burton

The longer-term goal remains to have safe autonomous systems with 'functional sufficiency'[lx] that the system can adapt to the changing world it finds itself in.

There was important early learning for the Programme, which remains true in 2022, that the challenges faced are even greater than anticipated.

> "Although there are some very impressive new commercial prototypes, it has become clear that the step from successful demo to 'prime time' is very significant."[lxi]

Tracking and demonstrating influence is problematic, and will require honest, and objective assessments by independents with the Programme's stakeholders and funders.

5 Emerging outcomes

This review has identified a number of emerging outcomes that the AAIP has contributed towards. It is not possible to ascribe a contribution to the AAIP compared to other factors that may also have supported the achievement of these outcomes, however, the review finds some correlation between the AAIP's investment in activities and these outcomes that have been thematically synthesised.

- **Improved awareness/ understanding/ cognition/ knowledge of the importance of safety assurance**

Whether as part of demonstrator projects, or through training, webinars or events, the AAIP's work is raising awareness and extending knowledge on how to assure RAS

- **Discovering and testing approaches with the potential to improve the efficiency of safety analysis / reducing uncertainty, risk and hazards**

The AAIP demonstrator model, used as intended, and improved over the course of this first iteration, is designed to contribute to one (or more) of the research blocks. All have safety embedded into their research, and can in turn influence the focus and design of future studies

- **Improved skills through direct application of knowledge 'in the wild' through demonstrator projects**

Demonstrator projects have built, tested or validated RAS in different contexts. The skills of those involved, based on case study evidence has increased, as evidenced by new approaches, ways of working, frameworks and formulating further research questions.

- **Improved assets that can benefit 'all' in future e.g., research, body of knowledge, datasets**

The legacy of the demonstrator projects has been both successor projects but also access to the findings and datasets generated and made publicly available.

- **Changed perceptions/attitudes**

As evidenced from the demonstrator project case study, a shared experience and different perspectives is contributing to a shift in perceptions and attitudes. A challenge for the future is how to support these changes at scale.

- **An ethos and approach that encourage frequent communication, collaboration and focus on safety**

York's AAIP team are seen to embody this culture. The 'deep collaborative bedrock' on which AAIP is built, and which is interwoven through all of its manifestations, builds and nurtures networks. The result is a growing international community and follow-on studies consistently funded.

- **A cross-domain, multi-disciplinary approach builds trust and stronger safety arguments**

Collaboration is key to the success of the Programme at all levels. The use of AMLAS and other guidance supports the questioning necessary to develop safety cases that are effective and are understood and used (as opposed to 'being in a ring binder on a shelf' in the words of one collaborator). AMLAS is thought to be accessible and useful to both AI designers, engineers as well as to safety experts – all of whom should be involved in considering safety from the earliest possible point of development.

- **Different questions being asked about safety of AI and machine learning that could improve manufacturer offers and client procurement i.e., capability to commission systems more safely**

Albeit anecdotal evidence at this stage, the confidence and authority from being part of a demonstrator project has led, in at least one case, to better procurement decisions. This outcome indicator (for example presented as money saved) could usefully be tracked through AAIP 2.

- **Increased confidence to deal with 'safety assurance in uncertainty'– willingness to take decisions**

Case study feedback from MSc module completers suggests that the training enables participants to feel up to date in their practice, and therefore, more confident in the decisions they take as safety engineers – and, potentially, more likely therefore to tackle complex safety challenges rather than not. The extent to which behaviours have already changed is as yet unknown – with more time required for the AAIP's guidance to be adopted and trailed.

- **Recognition of AI safety (particularly human factors) within workforce training**

Feedback from the NHS in particular shows that the Programme's emphasis on the importance of human factors is influencing training content.

- **Shared experiences across diverse sectors, occupations and roles without betraying commercial confidences**

The experience of Programme Fellows speaks of valuable cross-sector learning facilitated in a neutral space / space for public good. The fostering of relationships between members of the international community in turn may lead to different collaborations.

- **Projects that unite conversations between, and extend methods and approaches, with cross-sector applicability**

An intended outcome from demonstrator projects is learning and guidance that has the potential to be applicable cross-sector. The AAIP's domain agnostic position in turn promotes the conditions where learning can be shared across sectoral borders.

The international collaboration is supported the development of stronger guidance. The shared voice of the Programme, together with the opportunities afforded to share the AAIP's learnings is thought to be gaining greater exposure than comparable efforts by lobe actors. Case study feedback suggests collaborators are acting as advocates for the AAIP within their organisations.

- **Access to expertise that develop understanding and create influence**

A consistent finding across the case studies is the high value placed on the AAIP's team world leading expertise. This expertise engenders trust, and confidence and establishes influence. Creating influence and tracking its effects of course takes time.

- **AAIP is starting to contribute to narrowing the gaps, misalignments and lags between the advance in technology, workforce competence, industry and regulatory practice – a bridge**

The take up of training is lower than the AAIP team would have expected at this stage, and work remains to build both the blend of training and the staff to deliver it into AAIP 2.

Equipping the current and future workforce with the skills needed to confidently and safely manage and operate alongside RAS is arguably beyond the capacity of any one part of this landscape – and will require a co-ordinated approach across the international community. Nevertheless, the building blocks provided by the AAIP approach have demonstrated, at a small scale, the importance of training. Different stakeholders will be planning their workforce training strategies in different ways – and how the Programme seeks to influence these strategies

Outcomes that are observed or referenced by at least one case study participant include the following, however, caution should be applied to their generalisation in the absence of more extensive evidence.

- **Changed behaviours e.g., safer practices, more informed decision-making capability**
- **An AAIP community who will influence, persuade and change perceptions – which takes time.**

Many of these outcomes are made possible by a view helped by all case study participants (selected by the AAIP team) that the **University of York's reputation as world leaders in the area of safety is a key USP, underpinned by the continuity and levels of collaboration AAIP develops. This expertise and the AAIP's** continuity have been the glue that has bound different parts of the Programme together helping to realise some of the outcomes described.

Another way to visualise the outcomes achieved by the AAIP since 2018 is illustrated in the next table which considers outcomes experienced by different groups of stakeholders.

Logic Model by Stakeholder Group (based on case study evidence)

Regulators	Industry	Academia
• Improved awareness/ understanding/ cognition/ knowledge of the importance of safety • Changed perceptions/attitudes • Different questions being asked about safety of AI and machine learning that could improve manufacturer offers and client procurement i.e. capability to commission systems more safely • Increased confidence to deal with 'safety assurance in uncertainty'– willingness to take decisions • Recognition of AI safety (particularly human factors) within workforce training • Influence on (forthcoming) regulatory standards • Improved skills through direct application of knowledge in the context of demonstration projects	• Improved awareness/ understanding/ cognition/ knowledge of the importance of safety • Changed perceptions/attitudes • Different questions being asked about safety of AI and machine learning that could improve manufacturer offers and client procurement i.e. capability to commission systems more safely • Increased confidence to deal with 'safety assurance in uncertainty'– willingness to take decisions • Recognition of AI safety (particularly human factors) within workforce training • Changed behaviours e.g. safer practices, more informed decision-making capability • Improved skills through direct application of knowledge in the context of demonstrator projects • Increased ability to design for assurance	• Development of new assurance strategies and argue the safety of systems from 'first principles' • Improved awareness/ understanding/ cognition/ knowledge of the importance of safety • An AAIP community who will influence, persuade and change perceptions – which takes time • Improved assets that can benefit 'all' in future e.g. research, body of knowledge, datasets • Increased investment in further research • Improved skills through direct application of knowledge in the context of demonstrator projects • Discovering and testing approaches with the potential to improve the efficiency of safety analysis / reducing uncertainty, risk and hazard • Strengthened capacity to advance the safety assurance of autonomous systems • Expanding and sharing knowledge (as a provider of independent, practical guidance on the safety assurance of AS (2021) • Increased ability to design for assurance (through a research excellence approach)

6 Conclusions

How are industrial practices safer because of the AAIP's work?

The most tangible way that the AAIP has supported safer industrial practices is through the development of guidance supported by the underpinning research and practical experiences enjoyed by demonstrator project participants. AMLAS, as the most mature and first published guidance, can help individuals that work in a role that includes the assurance of safety to know what needs to be done differently in their organisation's practice to assure the safety of machine learning. It provides an approach that can support the development of a safety case, the first of which has been published and can hopefully inspire further case studies and practical applications by pioneers and early adopters in order that results can be generalisable in future.

With its adoption, at scale, over time, the AMLAS guidance has the potential to inform the improvement of safety practices across multiple domains across the world. Further plans to develop this, and other guidance and tools, should encourage the conditions whereby more industry partners use it in ways to develop robust safety cases, in turn building an evidence base for its practical application in real-world environments and contexts.

How has the AAIP impacted safety-critical sectors?

The AAIP team has, through their publicly funded research, made important progress on assurance issues, which will likely impact on safety critical sectors once the necessary levels of trust are in place for RAS implementation to become more widespread.

The maturity and progress in the healthcare sector is seen to be the most advanced, with forthcoming guidance that will be accessible across the clinician workforce.

The case studies and testimonials reviewed support the AAIP team's assertion that the 'demonstrator projects contribute evidenced, repeatable techniques for demonstrating the safety of autonomous systems.'[lxii] They have delivered and/or validated guidance that has fed into the Programme's Body of Knowledge.

The Programme's guidance, AMLAS being the most mature, has for the first time provided a way to include ML into a safety case. Those working in safety critical sectors have praised the guidance for it is technical content, but crucially, because it is designed with applicability in mind, i.e., it sets out how to do safety assurance (with helpful safety argument patterns incorporated) rather than just saying 'this needs to be done'.

The demonstrator projects have enabled outcomes for participants that have impacted their cognition, attitudes, behaviours and desire to continue collaborating on projects that further the capability to assure the safety of RAS in real-life environments. AAIP has also encouraged the conditions for collaborators to leverage other resources to further research and exploration catalysed or amplified by the demonstrators.

Utility of the domain-agnostic guidance arises from domain specific interpretation and application and requires sector specialists able to use the guidance and adapt it for their context. This approach works, recognising that regulatory constraints exist at domain-level.

How have we equipped safety engineers and others with the skills they need now?

The education and training developed by the Programme, informed and shaped by the research and real-life learning and guidance generated by the demonstrator projects, provides learners with both the methodologies and examples necessary to more confidently work on safety assurance and using the guidance, a common language to share and bring others along.

The gap between demonstrator and implementation remains significant. Consequently, the AAIP's strategy needs to be focussed not solely on meeting the current skills and knowledge needed by the current workforce, but also anticipate the future skills and knowledge of the future workforce. Moreover, the AAIP team report finding it hard to make the training offer exciting, but the market potential is just that, and the risk of missing a valuable opportunity is real. These challenges are addressed in the recommendations section of this report.

How to bridge the gap between the growing number of stakeholders who will benefit from training, and the team's capacity to deliver this to their required standards, will be a key challenge for AAIP 2, since the current targeted approach can only achieve scale and reach indirectly.

How have we guided the safe development of autonomous systems across the globe?

On the basis of the evidence reviewed and testimonials gathered, the AAIP can with greater confidence assert themselves to be a provider of independent, practical guidance on the safety assurance of autonomous vehicles.

The multidisciplinary collaborations made possible by the AAIP is one of its foremost assets; creating the conditions for complex challenges to be tackled, knowledge and connections to be extended, and the results to be more credible / accepted, compelling and widely shared and potentially applied than would otherwise be possible. Through embedding safety thinking into research, further collaboration, at different scales, means the opportunity for enduring change is created which realises all the potential that greater use of safe RAS could bring.

A small and select number of regulators have so far been involved and reported positively on how their experience and learning will help to shape the standards they plan to write. Regulators have valued the cross-domain applicability of the AAIP which means the Programme is well positioned to make a difference here. The first standards successes are already in evidence, but regulations change slowly, and the 'market' for regulator education is perhaps at an early stage.[lxiii] As with professional education and training more generally, a co-ordinated approach across the AAIP's partners will likely be required to make a noticeable impact.

As the pace of technological innovation continues to increase, a key area for the next phase of the AAIP will be to continue to work with regulators that define the safety parameters for industry to help them to understand 'assurance uncertainty' and how to structure and respond to this problem.

A challenge for all regulators in this space is how to develop 'agile regulation'[lxiv] and it is in this aspect in particular that the Programme has much to offer. Traditional debate and consensus models of agreeing standards will not keep pace with technological advances – and so limit the potential of the technology and dampen global competitiveness.

> *"There's a danger of regulation being too slow, or too restrictive, which would effectively mean that we couldn't exploit the full potential of the technology that might actually make things safer and better." (Professor Dr Simon Burton)*

The team have achieved greater penetration in some markets than in others, and while the international community is larger than in 2018, there remains much to do to grow and nurture this community.

Not explored in this review, but the AAIP team is aware of commercial organisations taking forward aspects of the AAIP approach. The ideal scenario is for these communities to join up wherever feasible to maximise the aggregate impact of learning across all technologies, sectors and jurisdictions.

7 Recommendations

This independent review, whilst not being an impact assessment has provided an opportunity to take stock of the emerging outcomes as a result of the investment in the AAIP, and its consequent deployment of energy towards various inter-related activities.

The consultants acknowledge the Programme's coherence of future activity around its five research pillars, and this is logical and developmental. In November 2022 the Programme reported that it wishes to build on achievements as follows:

1: Want to work more through partnerships and collaboration

- Partly enabled through facilities in ISA building

2: Application, validation and refinement of AAIP work

- With industry on development and assurance of RAS
- With regulators on shaping standards and regulatory frameworks

3: Undertaking new work, particularly support for growing usage of RAS

- Safety management systems
- Use of DevOps approaches and aligning safety and development

Staying ahead of the safety assurance curve will be vital to ensure the AAIP's market-leading status and credibility with due regard to some of the shared problems that exist:

"Working on deltas is something the whole ML community is working on and none of this is easy. We don't have all the answers yet." (AAIP collaboration workshop delegate, November 2022)

The Programme team already recognises a range of opportunities that might create the conditions for further future success in being able to catalyse the scale of safety assurance:

1: The use of the AAIP's new facilities – the Institute for Safety Autonomy – in York which provides bespoke indoor/outdoor laboratories for robotics software, hardware and advanced communications and opportunities for research, public outreach, industry and regulator collaboration.

2: Opportunities to apply for external funding with collaborators e.g., positioning the AAIP as an Artificial Intelligence hub and or part of a network of AI hubs.

In addition to these known trajectories for the AAIP, the consultants make the following recommendations as thought is given to the architecture of the Programme for its next chapter.

Recommendation 1:

Consider making the case for the investment in a comprehensive RAS safety assurance workforce development strategy and implementation plan.

This recommendation seeks to address the current deficit between AAIP capacity and mechanisms for educating the global market and the current and future projected need for skills and knowledge amongst the workforce across technologies, domains and jurisdictions.

This recommendation will force a great many actors to coalesce around this suggested priority area for action, domestically and internationally.

Consideration should be given to how this workstream is embedded not only into the AAIP Programme's capacity and plans for the future (i.e., additional tutor capacity / Research Fellow deployment), but also to how it can feature consistently in the work of localised centres within the international community that AAIP is helping to develop.

This recommendation has significant resource implications and it will be necessary to develop a costed proposal for strategy and implementation activities.

Recommendation 2:

In planning the next phase of AAIP investment and activity, it is recommended that a minimum of 2% and up to 3% of total Programme expenditure be dedicated to monitoring, evaluation and impact assessment activity in order to deepen the evidence base that can help demonstrate correlation between its work and the outcomes that emerge.

Routine monitoring aligned to a revised, agreed logic model, evaluation framework, key evaluation questions, indicators and data and evidence collection plan completed through a mix of internal AAIP and external 'learning partner' expertise will improve on the limitations of this snapshot review process conducted in 2022.

This approach should be considered for a minimum of at least five years from 2023 to 2028. It will support any ongoing funding opportunities, help leverage further assets towards the core mission of the Programme and provide an increasingly credible narrative amongst different stakeholders.

This approach would also be enabling formative and continuous Programme improvement in light of learning being captured from the community as well those that chose not to engage despite affecting, or potentially being affected, by the requirement to safely assure RAS.

End Notes

[i] Our team - Assuring Autonomy International Programme, University of York
[ii] Foresight review of robotics and autonomous systems. (lrfoundation.org.uk)
[iii] Source: Source: Grant application (G\100281). Page 13
[iv] AMLAS - Assuring Autonomy International Programme, University of York
[v] Please see report appendix for a copy of the case study guide
[vi] Assuring Autonomy International Programme, Executive Summary (extract from the funding application drawing on the Lloyd's Register Foundation Foresight review of robotics and autonomous systems (October 2021)
[vii] Foresight review of robotics and autonomous systems, serving a safer world, Lloyd's Register Foundation Report Series No 2016.1 (October 2026)
[viii] Ibid.
[ix] AMLAS - Assuring Autonomy International Programme, University of York
[x] Safety Assurance of autonomous systems in Complex Environments (SACE)
[xi] AMLAS - Assuring Autonomy International Programme, University of York
[xii] The full framework to provide safety assurance for manufacturers and adopters is scheduled for completion in February 2023
[xiii] Review of the AMLAS Methodology for Application in Healthcare
[xiv] Assurance of AI and Autonomous Systems: a Dstl biscuit book - GOV.UK (www.gov.uk)
[xv] SACE - Assuring Autonomy International Programme, University of York
[xvi] The Technical Cooperation Program - Wikipedia
[xvii] R. Hawkins, et al., Creating a safety assurance case for an ML satellite-based wildfire detection and alert system, arXiv preprint arXiv:2211.04530 (2022)
[xviii] Our team - Assuring Autonomy International Programme, University of York
[xix] British Standards Institution - Project (bsigroup.com)
[xx] https://www.iso.org/standard/83303.html
[xxi] Demonstrators - Assuring Autonomy International Programme, University of York November 2022
[xxii] how an AS explains its decisions
[xxiii] Leverage figure correct as of December 2022
[xxiv] Confident Safety Integration for Cobots
[xxv] Assuring the safety of cobots (CSI Cobot phase 1) - Assuring Autonomy International Programme, University of York
[xxvi] 12 journal papers, 25+ conference/workshop papers, 5 conference keynotes, tutorials and panels
[xxvii] £3.4m from the UKRI Trustworthy Autonomous Systems Programme, £360k from Dstl, UKAEA and £246k from EPSRC/Orca-Hub
[xxviii] IEEE Guide for Verification of Autonomous Systems, 4 CORE rank A conferences co-chaired and 3 journal special issues
[xxix] Explaining autonomous decisions - Assuring Autonomy International Programme, University of York
[xxx] For more details on the Commission's strategy for autonomous vehicles, please see: https://digital-strategy.ec.europa.eu/en/policies/connected-and-automated-mobility
[xxxi] AI in ambulance response - Assuring Autonomy International Programme, University of York
[xxxii] AI for Patient Consultations (corti.ai)
[xxxiii] The Systems Engineering Initiative for Patient Safety
[xxxiv] https://ori.ox.ac.uk/projects/rails/
[xxxv] Our team - Assuring Autonomy International Programme, University of York
[xxxvi] Shared control in autonomous driving - Assuring Autonomy International Programme, University of York
[xxxvii] Safe unmanned marine systems (ALADDIN) - Assuring Autonomy International Programme, University of York
[xxxviii] AAIP A Year in Review 2021, page 22
[xxxix] Total take up so far of the module is 22 (across 2 cohorts in 2021 and 2022)
[xl] The learning outcomes for the MSc can be found here: https://www.cs.york.ac.uk/professional/system-safety-engineering-courses/adts/

[xli] https://www.swansea.ac.uk/media/Remote-Control-and-Autonomous-Shipping-Final.pdf
[xlii] Staff from the VCA (Vehicle Certification Agency) have also been trained.
[xliii] https://www.york.ac.uk/assuring-autonomy/work-with-us/networks/.

[xliv] Safety-Driven Design of Machine Learning for Sepsis Treatment, Jia, Yan, Lawton, Tom, Burden, John et al (2 more authors) (2021). Journal of Biomedical Informatics. 103762 ISSN 1532-0464
[xlv] A suggestion which also aligns well with the EPSRC's direction of travel (see Chapter 5)
[xlvi] As an example, there is a part of Hamburg that is given over to driverless vehicles- https://www.moia.io/en/news-center/vwcv-moia-and-argo-ai-present-roadmap-for-autonomous-ride-pooling-in-hamburg
[xlvii] Source: *The AAIP Programme, reflections in September 2022*
[xlviii] Source: The Economic Impact of Technology Infrastructure for Advanced Robotics, NIST Economic Analysis Briefs 2, October 2016
[xlix] BEIS 'Made Smarter'. Review 2017.
[l] https://www.ukri.org/about-us/strategy-plans-and-data/. EPSRC's vision is for the UK to be recognised as the place where the most creative researchers can deliver world-leading engineering and physical sciences research
[li] https://www.ukri.org/opportunity/ukri-trustworthy-autonomous-systems-programme-responsibility/. In 2020, six projects, called nodes, are part of the UK Research and Innovation (UKRI) Trustworthy Autonomous Systems (TAS) programme, and will undertake fundamental, creative and multidisciplinary research in various areas key to ensure autonomous systems can be built in a way society can trust and use.
[lii] 'The TAS Node in Resilience project is a 30-month project which brings together the disciplines of computer science, engineering, law, mathematics, philosophy and psychology from five UK Universities, to develop a comprehensive toolbox of principles, methods, and systematic approaches for the engineering of resilient autonomous systems and systems of systems'
[liii] Going forward, all demonstrator projects should align to at least one of the research blocks or pillars developed by the Programme – each of which will include domain agnostic guidance. (1) Assurance of machine learning in autonomous systems (AMLAS), 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) and 5) Societal acceptability of autonomous systems (SOCA). https://www.york.ac.uk/assuring-autonomy/research/research/
[liv] National AI Strategy. While the focus of this first strategy is about safety and security of citizens, it does includes the recommendations to 'Establish medium and long term horizon scanning functions to increase government's awareness of AI safety', and 'Work with The Alan Turing Institute to update guidance on AI ethics and safety in the public sector'
[lv] They are currently developing plans which align with the investments outlined within the EPSRC Delivery plan 2022-25
[lvi] Source: World Robotics 2020 Industrial Robots. International Federation of Robotics. In 2019, 4.8% (18,000 out of more than 373,000) industrial robots installed were cobots, an increase of 11% compared to 2018
[lvii] Source: 'The Collaborative Robot Market. Interact analysis (2019)
[lviii] Source: 'Assuring Autonomy International Programme. A Year in Review' 2019
[lix] A review in May 2020 identified that the community was "Geographically dispersed - mainly across the UK, USA, Australia, Germany and to a lesser extent Japan, other parts of Europe, UAE and Canada".
[lx] From a blog by Dr Burton, 'Staying safe in an uncertain world. Assurance strategies for automated driving systems' (June 2020).
[lxi] Source: 'Assuring Autonomy International Programme. A Year in Review' 2018
[lxii] AAIP A Year in Review 2021, page 22
[lxiii] In November 2022, the Programme reported that it is now running continuing professional development (CPD) courses for the Marine and Coastguard Agency (MCA)
[lxiv] See also NESTA's work on anticipatory regulation which feels closely related. https://www.nesta.org.uk/feature/innovation-methods/anticipatory-regulation/?gclid=Cj0KCQiAg_KbBhDLARIsANx7wAyRt03ZWZTk6ffks1g2323Dtz3FUvxYRPPNIuU-nUtGt95FpVxmMVsaAvOkEALw_wcB