UKRI
Trustworthy
Autonomous
Systems Hub

TAS Thought Pieces

January 2023

FOREWORD

Over the last 2 years, the Trustworthy Autonomous Systems (TAS) Programme has involved over 30 institutions and more than 600 researchers from different disciplines and with different perspectives on trust. We work closely with industry, government, and the public to ensure our research will be socially beneficial to all key stakeholders and the wider society. In the TAS Hub alone, we launched over 40 projects touching on sectors such as healthcare, autonomous vehicles, and creative industries working with over 100 industry partners.

Given the breadth of the research landscape TAS covers, it is not always easy to understand what TAS will deliver for a particular industry or sector and how we, as a community, are working together and combining our approaches to solve the most pressing challenges. It is therefore important to take a step back and reflect, as a community, on what we have achieved so far, what are the upcoming challenges, and how we are going about addressing them. Presenting such reflections and plans to our industrial partners and government will ensure we align our objectives and, we hope, help them identify the best approaches to answer their most pressing questions.

To address this need, we launched a series of workshops with over 40 members of the TAS community from the Hub and the Nodes. The output of these workshops is this collection of thought leadership articles. Each article focuses on a specific aspect of trust: Functionality, Trust, Security, Resilience, Verifiability, Governance and Regulation. The collection brings together the views from across the TAS programme on key challenges, approaches, and early research outputs. We worked with researchers from different backgrounds, disciplines, and institutions to ensure this collection is representative of the diversity of perspectives that exist across the TAS research community.

I would like to thank all the researchers for their time and effort in participating in the workshops and reviewing the resulting material. I am also grateful for the support from Thales and their contributions to the workshops and some of the use-cases presented here.

I hope this collection of thought leadership articles will serve as a first step for new conversations with the TAS community and helps grow our collaborations with industry, government, and the public.

Professor Sarvapali (Gopal) Ramchurn

Director, UKRI Trustworthy Autonomous Systems Hub

UKRI Trustworthy Autonomous Systems Hub

FUNCTIONALITY

INTELLIGENT ROBOTS AND SELF-GOVERNING SWARMS: THE FUTURE OF FUNCTIONALITY IN AUTONOMOUS SYSTEMS

Developing technology that we can trust and rely on is key to its future acceptance and expansion within our society. Over recent decades we have been developing increasingly complex systems to carry out specified tasks and operate in a pre-determined way; their functionality, largely controlled by us.

However, as our knowledge of technology advances, so does the capability of the technology itself. Autonomous systems (AS) are emerging with the tantalising potential to develop and adapt functionality for themselves, without human input. These are exciting advances – but not without their challenges. How will this autonomy impact on how they operate in complex situations in the real world? How can we realistically predict what they are going to do and ensure they always make the right decisions? And how does this self-learning impact on our trust?

These mission-critical questions are at the heart of the Functionality Node of the UKRI Trustworthy Autonomous Systems (TAS) Programme – a £33m multi-disciplinary research programme funded as part of the Strategic Priorities Fund comprising six Nodes – separate research projects examining individual aspects of trust in autonomous systems.

A number of key challenges are being examined: how does giving autonomous systems the ability to evolve their functionality influence how we specify, design, verify, regulate and build trust in these systems? How can we monitor and check how they are operating in unpredictable environments? Who is to blame if something goes wrong? There are complex considerations around expectation, responsibility and ethics.

Dr Shane Windsor from the University of Bristol sums up the ultimate goal: “In our work, we are interested in systems with functionality that evolves through time from emergent behaviours, that are safe, reliable, resilient, ethical and trustworthy”.

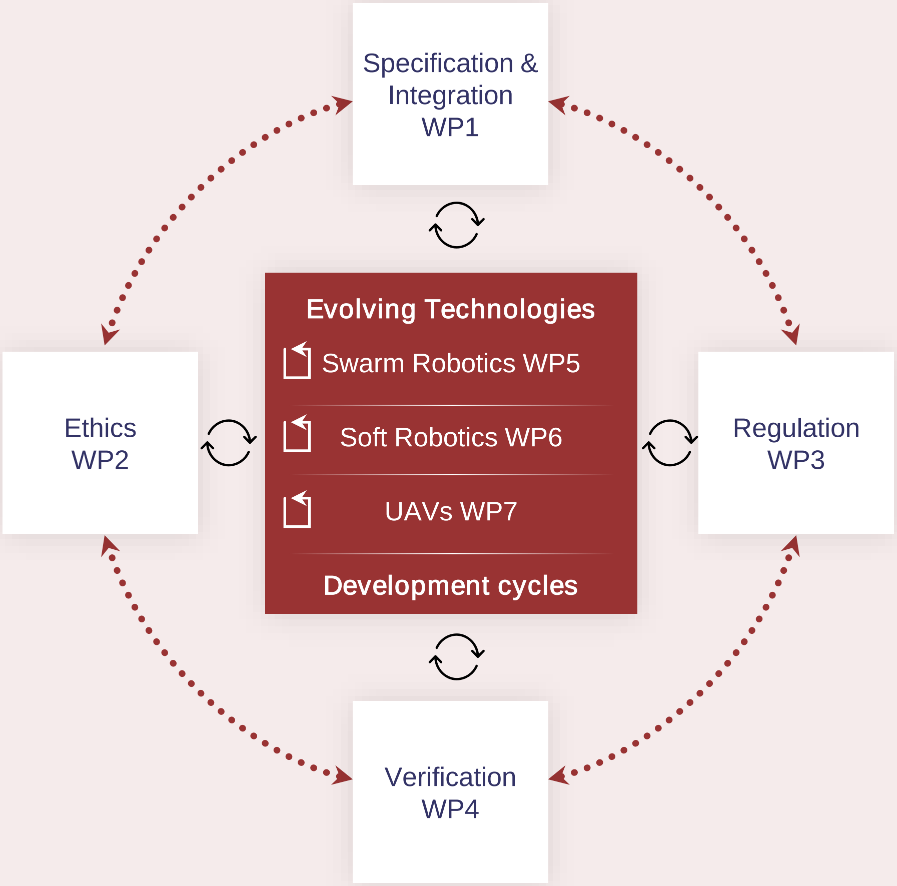
DESIGN FOR LIFE

The starting point for addressing some of these challenges is at the development stage.

Researchers are examining if changes need to be made in the design of autonomous systems to assist with evolving functionality once they are in operation. Professor Kerstin Eder from the University of Bristol emphasises the importance of this:

“There are challenging research questions around specifying for evolution and adaptation, as autonomous systems adapt over time. We need to identify design principles and operational techniques that enable trustworthy evolving functionality.”

A Design-for-Trustworthiness framework for adaptive autonomous systems is being developed which will help create guidelines, methodologies and technologies for evolving functionality. There are four focused research themes - specification, verification, ethics and regulation – and three adaptive technology development use cases: swarm robotics, soft robotics and unmanned aerial vehicles.



Overview of the TAS Functionality Node structure, 4 design-for-trustworthiness process research themes (WP1-4) around 3 developing technology use cases (WP5-7) (from Windsor et al., 2022).

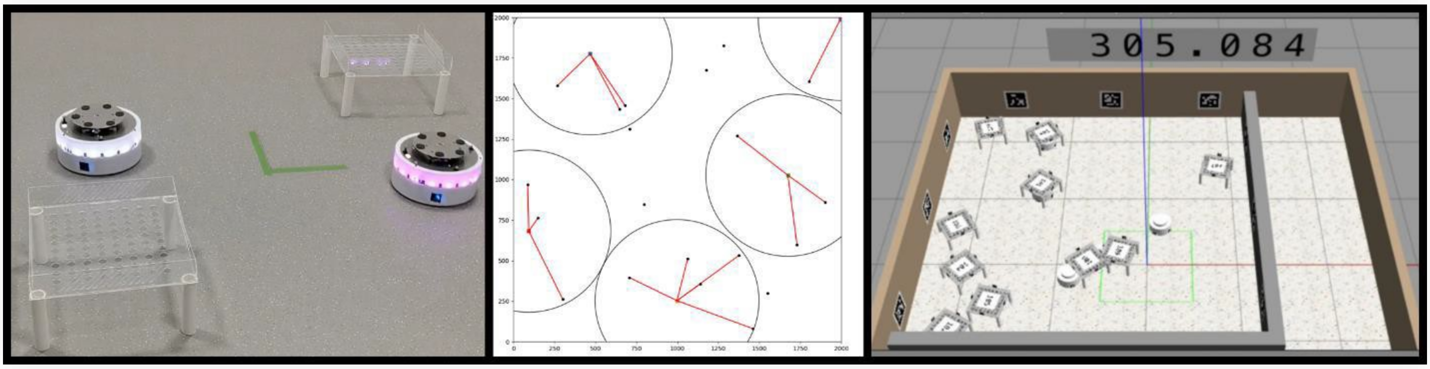
THE SWARM EFFECT

One of the key areas of research in furthering our understanding of adaptive functionality is observing how robots work together, as well as individually.

It is still early days for swarm technology - groups of robots operating together to achieve an objective - but there is significant interest in its potential. It is hoped that robotic swarms, with increasing levels of autonomy, will be able to adapt to their environment, enabling them to be widely used in many critical situations; emergency rescue, healthcare and logistics, for example.

The signs so far are positive, but translating this potential from theory into real life scenarios is far from easy. How do we ensure that what works in a test environment works in practice? Swarm functionalities, such as information-gathering, decision-making and sychronisation, need to be measurable, reliable, ethical, resilient and safe. What happens if one autonomous element is swapped or upgraded – does this impact the safety and functionality of the entire swarm?

TAS research teams are developing the cyber-physical infrastructure for swarms, including low and high-fidelity simulators and physical test beds. The TAS Functionality Node is studying swarm solutions for storage and retrieval in unstructured environments like cloakrooms. Low-fidelity simulators allow high-level ideas and concepts to be tested and explored quickly with standard computers. Slower, high-fidelity simulators provide the next step in gathering accurate results and highlighting potential issues. Swarm robot hardware, such as the DOTS platform, enables real-world testing and assessment of user and public trust in these systems.



The TAS Functionality-Toshiba DOTS cloakroom attendant robot (left), the low fidelity 2D simulator (middle) and the high-fidelity 3D simulator (from Windsor et al., 2022).

SWARM RESEARCH - THINKING ‘OUT OF THE BOX’

There is a growing appetite for ‘out-of-the-box’ swarm robotics - systems that can be used by operators with limited expertise and training. Work is underway into whether swarms could be deployed without complex set-up or infrastructure. Real-life simulation (or as close to it as possible) is key to exploring this ‘out-of-the-box’ potential.

A test facility for autonomous systems has been set up in Bristol where the TAS Functionality Node is based. The Bristol Robotics Lab is an open-access testbed for swarm robotic experimentation. Featuring a swarm of up to twenty robots that can be observed by publicly accessible webcams, the aim of the Lab is to model and ‘play’ with swarm AS scenarios in a more ‘out-of-the-box’ way.

It is hoped that the findings can help address challenges around perceptions of trust and responsibility, which could lead to exciting positive impacts on our lives. Dr Sabine Hauert from the TAS Functionality Node at the University of Bristol says there is enormous potential:

“For me, the most exciting areas of application are those that interact with the real world and human beings. Particularly, the idea of robots for intra-logistics could be useful for the third-party sector, local communities and the local sector circular economy. Could ‘out-of-the-box’ swarms be used on small scales for bakeries, foodbanks, small shops and care homes?”

Swarm arena for testing functionality of the DOTS at the Bristol Robotics Laboratory

[1] https://www.youtube.com/watch?v=Vwg7e-W7KZw

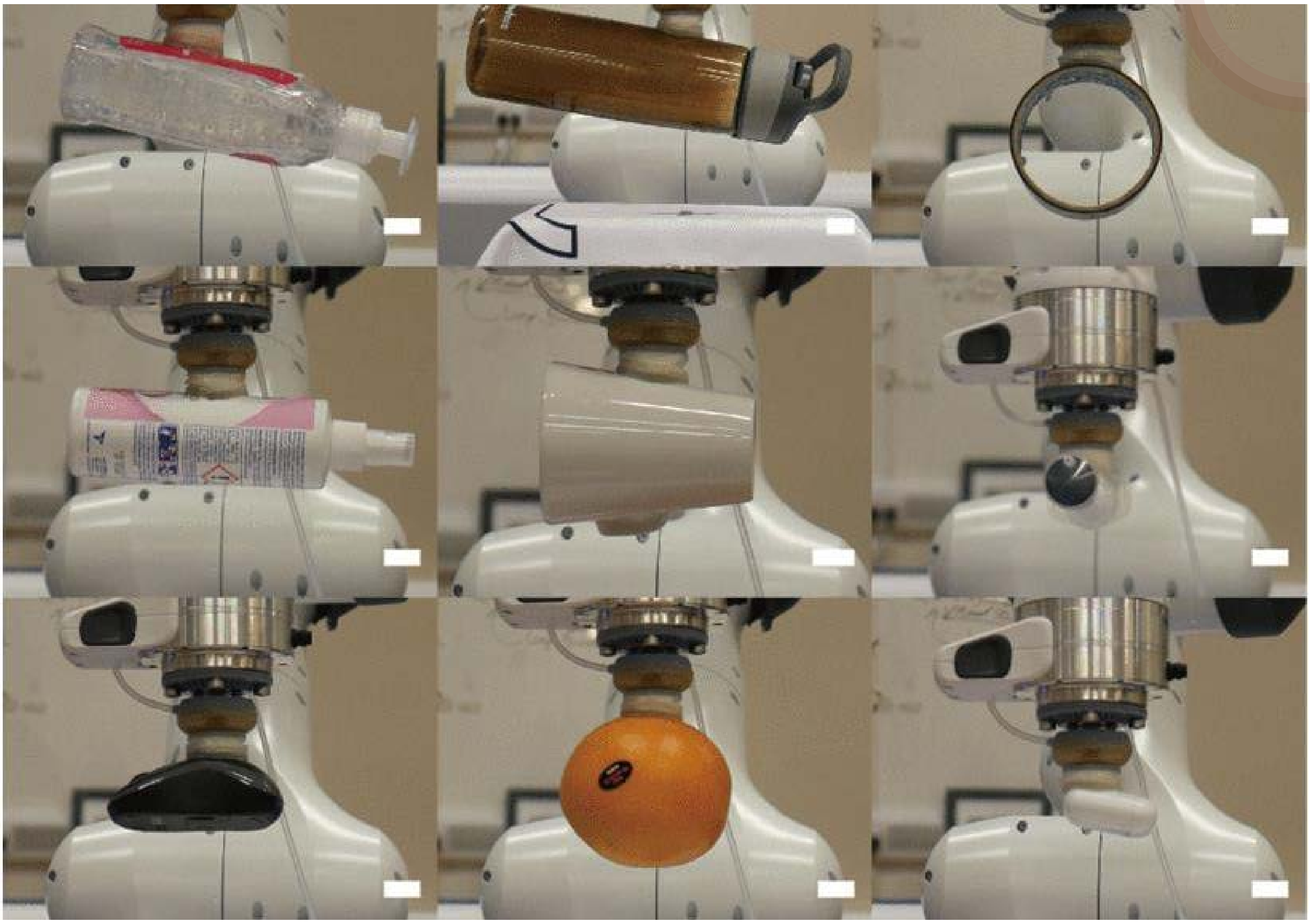
TAS research is also looking at better communication with swarms. The TAS Functionality Node has developed a web app enabling individuals to interact with a virtual swarm, modifying behaviours through performance and control metrics, and assessing real time trade-offs without the need for programming knowledge.

However, as Dr Sabine Hauert emphasises, test environments can only go so far and more real-world data is needed for future advancements:

"The thing we are missing is more Living Labs. Places in the community that we can go to test some of these systems in a more meaningful way in controlled real-world settings. This would open up applications in construction, environmental monitoring and healthcare."

SOFT ROBOTICS

Another key area of research around adaptive functionality is soft robotics. These are robots made from non-traditional materials that are – as the name suggests – soft and flexible. With these robots, the materials' properties do some of the work, rather than the need for each element to be controlled using motors and mechanics.



An adaptive, contact triggered, soft suction cup on a 7 degree of freedom, Franka Emika, robotic arm picking up objects of different shapes, curvatures, and textures (from Yue et al., 2022)

TAS researchers are currently working on making the functionality of soft robotics more predictable by using modular components. Guaranteeing the performance of a single module can generate more predictable outcomes when it interacts with other modules.

In a similar way to swarms, advances in this area would open up a wealth of applications for wider society. TAS researchers are, for example, studying multifunctional grippers for manufacturing that would limit downtime and aid productivity; robots for picking and packing; and in the medical field, soft robotic surgical assist tools.

DRONES

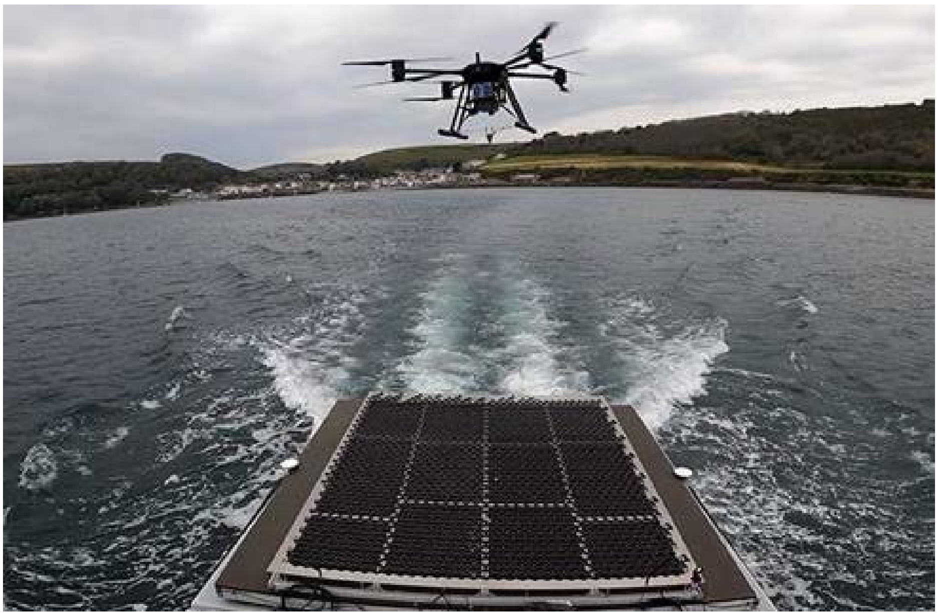
Progress is also being made in adaptive functionality within the field of unmanned aerial vehicles (UAVs) – more commonly knowns as drones.

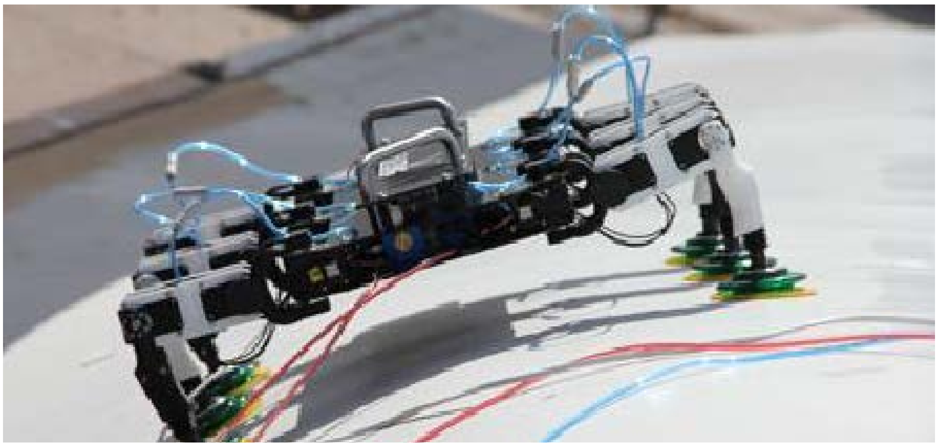
Researchers are looking into the use of machine learning for flight control of UAVs. Here, a drone would adapt as it flies along, learning from experience of different conditions and environments, such as rural or complex urban environments, via sensory feedback. As with soft robotics, UAV research is bio-inspired, aiming for the manoeuvrability and adaptability that we see in animals and birds. This opens up significant potential in infrastructure monitoring, surveillance, emergency response, logistics and even personal transport and aviation.

Thales are exploring the use of UAVs in offshore wind farm inspection, maritime search and rescue and ground vehicle resupply. The aim is to further expand potential uses of self-learning autonomous systems. Dr Matt Ball, Chief Scientist from Thales UK, says it is compelling research:

“We are working in partnership with academia to learn the general principles and results that we can apply. For example, in a project combining different autonomous systems on an offshore windfarm, we are using an uncrewed surface vehicle to get out to the windfarm, and then a UAV deploying a crawler robot for inspection and repair of the blade.”

[2] From: https://s.wsj.net/public/resources/images/TE-AB752_SOFTRO_M_20180307180915.jpg





Images from multi-platform robotic inspection system windfarm trials as part of the Offshore Renewable Energy (ORE) Catapult MIMRee project involving both Thales and University of Bristol.

Another specific scenario is enabling quadcopters to carry parcels in urban areas using reinforcement learning. The machine learning is initially being tested in a simulation environment and will then be expanded into a scaled test urban environment.

Studies are also underway into ground and airspace risk modelling and the future of crowded shared airspace.

FUTURE IMPACTS OF ADAPTIVE FUNCTIONALITY

Two key questions remain at the heart of evolving technology: how will autonomy make a difference - and what will we, as society, accept?

Ambitions surrounding autonomous systems and artificial intelligence span every part of our lives – macro to micro. From emergency rescues to clinical care; swarms of UAVs in disaster recovery scenarios to wearable robotics controlling cellular level swarms for wound healing.

Dr Shane Windsor says the goal is to create systems that can adapt to different locations and environments, and ultimately make our lives easier:

“I would love to see robots in the everyday world. Adaptation is a key enabler. You can design a static solution with current techniques, but it will only work for a limited set of situations. But an adaptive AS that can cope with high levels of difference and suit the environment they are working in, that is where I would like to see things going.”

The focus now is on taking the technological advances that we are seeing in a test environment and transitioning them into the real world – and knowing how far and how fast to go with this, as Dr Shane Windsor explains:

“There are a lot of things we can do currently and are technically on the horizon, but there will be a difference between what we can do and what we do - for example due to security and regulation - as well as questions about whether these are the right things to be doing.”

The promise is that evolving autonomous systems will add value to our lives. They can take us out of extreme situations, prevent us from risking our lives in dangerous environments, help us with our daily logistics, advance our medical abilities and assist with our care needs.

But this promise depends on trust.

Public perceptions are complex and there has always been a degree of caution about what ‘intelligent’ technology may mean for our way of life. In order to trust, we need our evolving autonomous systems to be reliable, human-centric and ethical. Ongoing research into adaptive functionality is key to this. We need to be totally on board for the ride, actively pushing for this potentially life-changing technology and not just passively accepting it as our future.

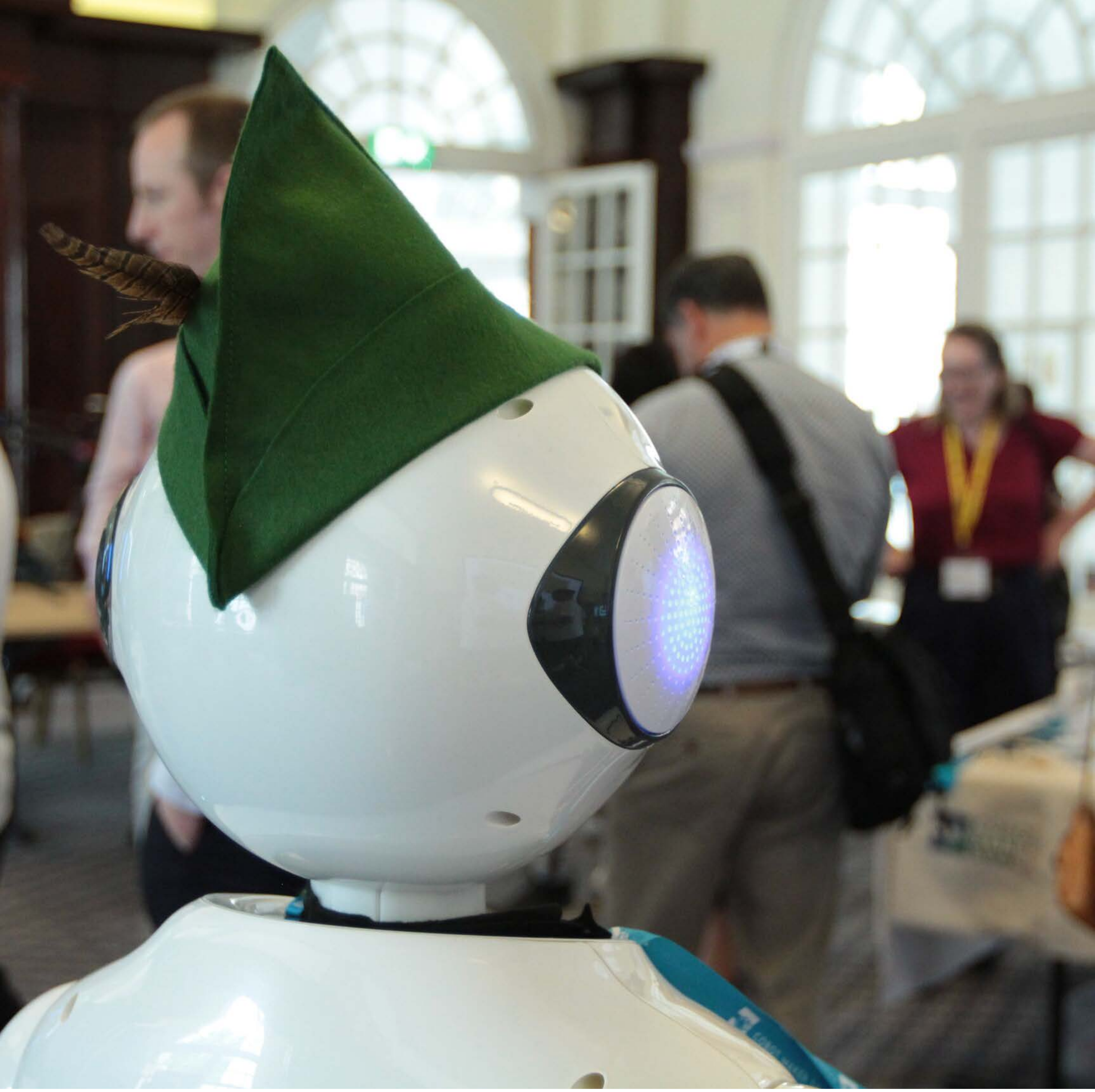
REFERENCES

Windsor, S., Downer, J., Eder, K., Hauert, S., Ives., J., Rossiter, 2021. UKRI Trustworthy Autonomous Systems Node on Functionality Annual Report- 2020-2021. Please contact shane.windsor@bristol.ac.uk for further information.

Windsor, S., 2021. UKRI TAS Node in Functionality: Research Talk: [accessed 06/05/22] https://vimeo.com/581077906

Yue, T., Si, W., Partridge, A. J., Yang, C., Conn, A. T., Bloomfield-Gadêlha, H., Rossiter, J. M., 2022. A Contact-triggered Adaptive Soft Suction Cup. IEEE Robotics and Automation Letters, 7(2), 3600 - 3607. https://doi.org/10.1109/LRA.2022.3147245

Web Ref 2 - Thales MIMRee project windfarm trials: [accessed 06/05/22] https://www.thalesgroup.com/en/united-kingdom/news/meet-robot-team-will-be-vital-future- offshore-wind-and-net-zero



UKRI Trustworthy Autonomous Systems Hub

GOVERNANCE AND REGULATION

REGULATING THE ROBOTS – THE CHALLENGES OF GOVERNING EVER-SMARTER TECHNOLOGY

As Autonomous Systems (AS) and Artificial Intelligence (AI) take significant steps forward in capabilities, a number of challenging questions emerge. How do we ensure that our intelligent machines continually meet certain standards? How can the law keep pace with machine learning and its diverse applications? Where does human input end – and who or what is responsible if something goes wrong?

These complex issues are among those being addressed by the UKRI Trustworthy Autonomous Systems (TAS) Programme – a £33m multi-disciplinary research programme funded as part of the Strategic Priorities Fund. Governance and Regulation makes up one of the six TAS Nodes – focused research projects examining individual aspects of trust in autonomous systems. It was also the first topic to be examined in a series of multi-disciplinary workshops, exploring the complexities and looking at what the future may hold.

The world of AS and AI is fast-moving and difficult to govern. How, then, do we begin assessing what governance and regulation is needed? And once that is in place, how do we enforce and monitor it?

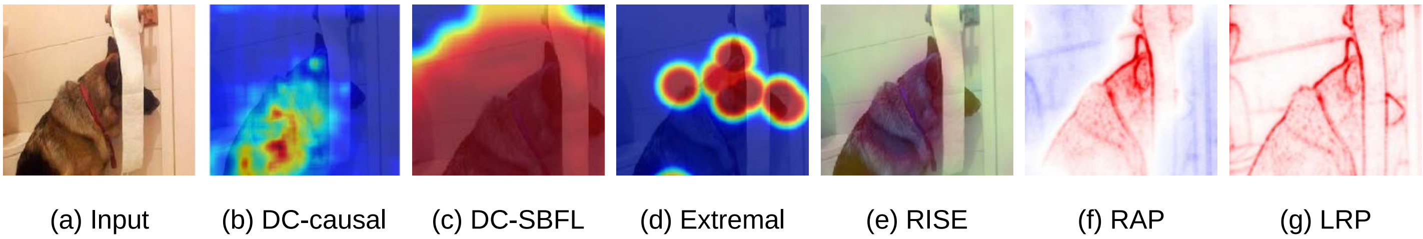
A range of methodologies and early results are being examined by the TAS programme, broadly within the parameters of four main areas of research: legal and social implications; machine learning, data and ethics; formal testing methods and implications in the design process. The aim is to create a Responsibility Framework which can inform and assist the regulators, backed up by real-world examples.

How to govern an AI world

Designing and deploying autonomous systems requires a multi-pronged approach. The use of standards in the design phase requires agreement on what those standards are. Once the systems are in use, new issues arise: human input is unpredictable, as is human interpretation of information. Where does our input end and machine learning begin? We need the real-world data from the autonomous systems in action, but how do we interpret this, and how can we use that data to help shape future governance?

TAS researchers have been working with key partners on a range of studies in different fields to examine this. Medical devices and AI analysis are one of the most fast-growing and important applications of TAS and Governance and Regulation research.

One example is tumour detection. Machine learning can be used to review MRI images of the brain, looking for patterns that might indicate tumour growth. Evaluation of this information, combined with other methods such as heatmaps, has the capacity to greatly advance technological capabilities and is an exciting glimpse into the future of medical testing.

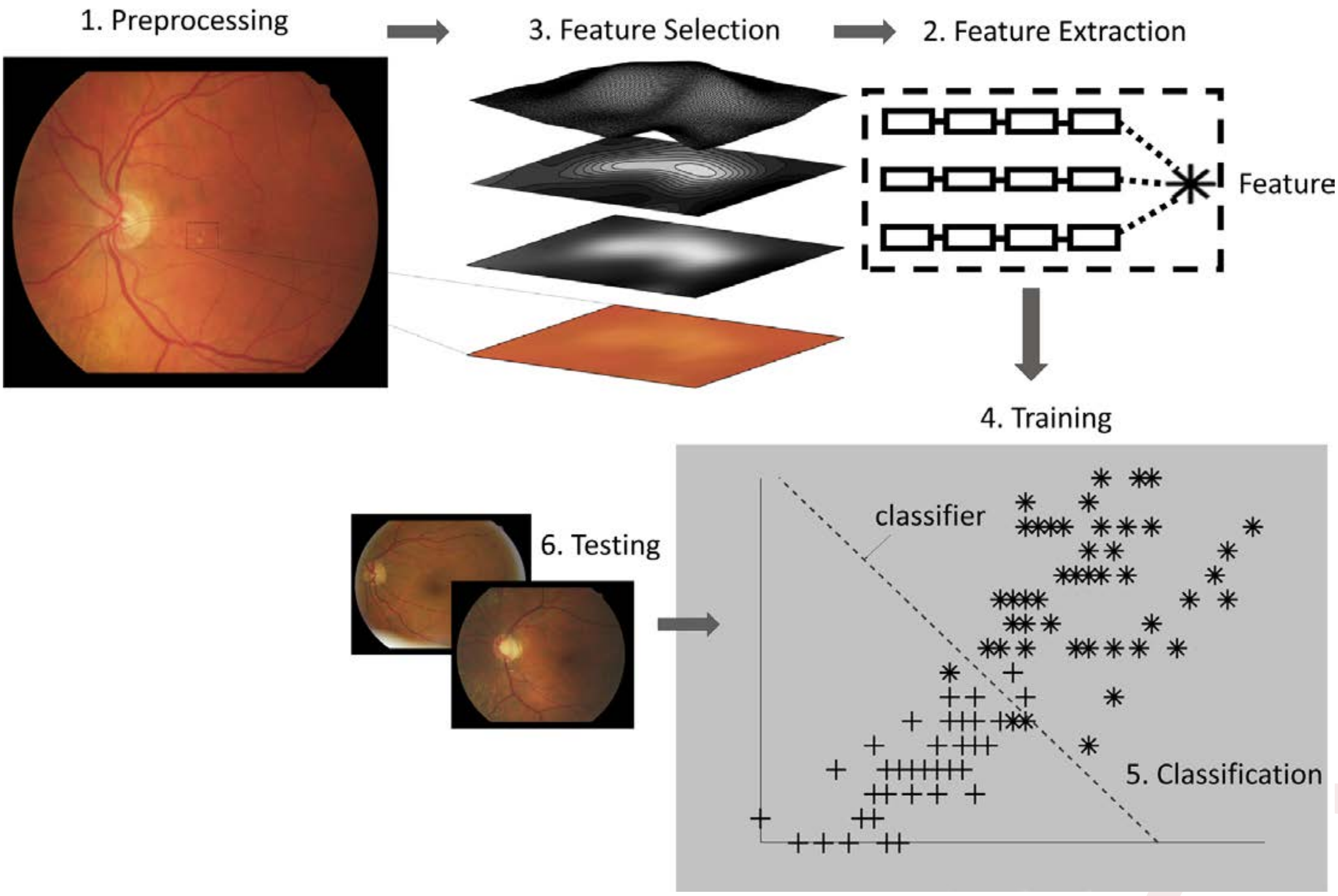


(a) Input (b) DC-causal (c) DC-SBFL (d) Extremal (e) RISE (f) RAP (g) LRP

Explaining the classification of a partially occluded image of a German Shepard dog (a), DC-Causal (b) performs best of the classification explanation tools (from Chockler et al., 2021)

Elsewhere, researchers are in contact with NHS Scotland in relation to the screening process for Human Papilloma Virus (HPV), a key indicator for development of cervical cancer. Autonomous systems and devices are also being used to help research colorectal cancer and explain the gut microbiome.

In the field of optometry, researchers have been looking into how AI can help mitigate human errors and improve diagnosis of eye disease, including early age-related macular degeneration. The use of machine learning works particularly well in helping understand decisions and errors in human-agent teams. As Dr Hana Chockler of King's College London says, this is exciting work in progress: "We're not yet treating patients, but the autonomous system and the machine learning components help in research, and their prediction seems to be a very large part of causality."



Schematic of a supervised machine learning pipeline for age-related macular degeneration in Optometry (from Pead et al., 2019)

The world of transport poses new questions regarding responsibility and regulation. How can AI account for and adapt to discrepancies in human behaviour and unpredictable real-world scenarios? For autonomous vehicles (AVs) to function correctly and compliantly, they must make crucial decisions on fast-changing data that comply with existing traffic laws. Such situations, however, are not always clear-cut. Likewise, laws themselves contain many discrepancies, real-world complexities and contradictions and are widely open to human interpretation.

Prof Burkhard Schafer from the University of Edinburgh explains: “How can we or should we at all represent these things, such as a near miss? We let humans get away with it. We might not want to let machines get away with it because machines are a different type of accuracy for norms than a human driver as we make excuses for human drivers that we do not necessarily want to make for a machine.”

Another aspect of autonomous vehicles is that of licensing. If cars were fully autonomous, would you need a driving licence? Not needing a licence seems highly unlikely. Even in systems where automation has been achieved to a very high extent, such as trains, there is still typically a human operator ready to take over if necessary. It may be that we may need to move to a very different type of a driving licence – perhaps aligned with what currently exists with automatic and manual licences - and one that requires updating as new automated features are introduced.

1 Ensure that CAVs reduce physical harm to person.

2 Prevent unsafe use by inherently safe design.

3 Define clear standards for responsible open road testing.

4 Consider revision of traffic rules to promote safety of CAVs and investigate exceptions to non-compliance with existing rules by CAVs.

5 Redress inequalities in vulnerability among roads users.

6 Mange dilemmas by principles of risk distribution and shared ethical principles.

7 Safeguard informational privacy and informed consent.

8 Enable user choice, seek informed consent options and develop related best practice industry standards.

9 Develop measures to foster protection of individuals at group level.

10 Develop transparency strategies to inform users and pedestrians about data collection and associated rights.

11 Prevent discriminatory differential service provision.

12 Audit CAV algorithms

13 Identify and protect CAV relevant high-values datasets as public and open infrastructural resource.

14 Reduce opacity in algorithmic decisions.

15 Reduce opacity in algorithmic decisions.

16 Identify the obligations of different agents involved in CAVs.

17 Promote a culture of responsibility with respect to the obligations associated with CAVs.

18 Ensure accountability for the behavior of CAVs (duty to explain).

19 Promote a fair system for the attribution of moral and legal culpability for the behavior of CAVs.

20 Create fair and effective mechanism for granting compensation to victims of crashes or other accidents involving CAVs.

Twenty principles for the ethics of connected autonomous vehicles to support researchers, policymakers, manufacturers and deployers (from European Commission, 2020)

MITIGATING RISK AND FAILURE

Another area of complexity arises from when things go wrong. Autonomous systems force us to examine how we might restore trust in a system after there’s been a problem or a failure. Where does the responsibility lie? Where does the apology come from if something goes wrong? Insuring against risk is complex and can have unintended consequences. If tough rules and stringent insurance regulations are imposed universally, for example, would this deter smaller companies from participating in developing the technology? And would imposing high insurance premiums price some innovators out the market, thereby limiting growth?

According to Prof Ramamoorthy from the University of Edinburgh: “We often think about insurance just in terms of safety; however, insurance plays an important role in the development of the sector. If you're a small company, trying to deploy a new product with these kinds of failure modes, there's simply no way unless there's some way of ensuring the company itself.”

Insurance companies are important in shaping regulation from the legal perspective. But to quantify risk, companies need data. Part of the insurance business model is to insure human skills; exactly how much risk can it take?

Prof Burkhard Schafer highlights the issue: “The moment you have an event that affects all [autonomous] cars because they're all using the same software and it fails at the same time. If that is your risk model, then our insurance companies are massively undercapitalized to cover that sort of thing.”

THE COMPLEXITIES OF INTERNATIONAL REGULATION

Cross-border legal questions surrounding AI are far from clear cut. Different countries have different approaches to governance and jurisdiction, so a single set of regulations is an impossibility. The EU have published a proposal for regulation on AI, which will affect most autonomous systems under development. It is being examined in depth by TAS who conclude that the idea of a multi-national regulatory approach is ambitious, but a good one. However, overall, it has significant shortcomings - in particular with potential dangers to the rule of law and democracy. Who can enforce the regulations? Does it give too much power to technology-setting bodies with little judicial control?

In a post-Brexit UK, we would also encounter some issues reminiscent of GDPR changes. The EU AI Act would have some overseas reach, and with a similar US initiative being announced, the UK risks falling between two large regulatory blocks.

These questions open the door for TAS researchers to work closer with standard-setting bodies to ensure that ethical and legal considerations are properly reflected in these standards.

Trust and the law

There are also challenges around how we create laws that are flexible enough to temporarily respond to an emergency without undermining trust in the technology and the legal system itself.

During the COVID pandemic, for example, there was a need for a rapid technological solution to a huge social problem – in essence, the rollout of the COVID-19 Track and Trace system. The laws on data permissions were temporarily overwritten to ensure the system could be implemented more quickly.

Prof Burkhard Schafer said the research drew some interesting conclusions: “We looked into responses to emergency and crisis and one of the findings is that some safeguards were unnecessary, especially in the UK context, where Parliament can't bind [devolved] Parliament and there is no constitutional court to necessarily set things back after the crisis is over.”

There are legislative techniques that can mitigate this, such as sunset clauses – provision that sets an expiry date when part of a law will cease to have effect. However, this poses technical challenges for the design of autonomous systems. For example, they may have to be designed with ‘contestability’ and ‘temporality’ as features, just in case legislators want to change the applicable law after they are deployed.

THE FUTURE OF AUTONOMOUS SYSTEMS

What next, then, for the regulation and governance of autonomous systems and AI? Some of the key findings of the Governance and Regulation workshop prove interesting reading.

One of the most notable conclusions relates to the difference between ideology and what is actually achievable in practice. According to Prof Ram Ramamoorthy: “There is a big gap between what we wish we could say about AS and what we can actually say now, and we may never achieve the security level of robustness alluded to.”

There is still much work needed in the areas of ethics, responsibility and machine learning. Do we need separate regulations for autonomous systems in isolation, as opposed to when humans and AS are both instrumental in their operation?

Dr Hana Chockler highlights the issue: “Physicians are getting acquainted with machine learning systems, showing them where the tumour is or what are the markers, but what if the physician is even more hands off and this is the norm. For example, directing the autonomous system to perform the surgery- a completely different strategy will be needed.”

In the end, it's all about mitigating risk, and the reality is that governance and regulation has to be carefully tailored and adapted as technology advances – and one size most definitely does not fit all.

REFERENCES

Chockler, H., Kroening, D., Sun, Y., 2021. Explanations for Occluded Images. in Proceedings of International Conference on Computer Vision (ICCV).

https://arxiv.org/pdf/2103.03622.pdf

Pead, E., Megaw, R. Cameron, J., Fleming, A., Dhillon, B., Trucco, E., MacGillivray, T., 2019. Automated detection of age-related macular degeneration in color fundus photography: a systematic review. Survey of Ophthalmology. Vol. 64, p. 498-551.

https://doi.org/10.1016/j.survophthal.2019.02.003

European Commission, Directorate-General for Research and Innovation, 2020. Ethics of connected and automated vehicles, Publications Office,

https://doi.org/10.1016/j.survophthal.2019.02.003

About the Trustworthy Autonomous Systems (TAS) Hub

The TAS Hub sits at the centre of the £33M Trustworthy Autonomous Systems Programme, funded by the UKRI Strategic Priorities Fund. Its role is to coordinate and work with six research nodes to establish a collaborative platform for the UK to enable the development of socially beneficial autonomous systems that are both trustworthy in principle and trusted in practice by individuals, society and government. For more information please visit the website:

www.tas.ac.uk

UKRI Trustworthy Autonomous Systems Hub

RESILIENCE

DEVELOPING RESILIENT AUTONOMOUS SYSTEMS WE CAN COUNT ON

Reliability and predictability are important in our lives. We like to be sure that when we switch on our smartphone, we can make calls and access our apps, or when we turn on the ignition, our car will start and we can drive to our destination. We rely on our machines to behave as we expect them to. We need to trust them – and for this we need them to be resilient.

Resilience is crucial to the development of ever-smarter technology and plays an essential role in ensuring that our autonomous systems are reliable and trustworthy. These complex issues are among those being addressed by the UKRI Trustworthy Autonomous Systems (TAS) Programme – a £33m multi-disciplinary research programme funded as part of the UKRI Strategic Priorities Fund. Resilience makes up one of the six TAS Nodes – focused research projects examining individual aspects of trust in autonomous systems (AS) through new research.

What does resilience in Autonomous Systems look like?

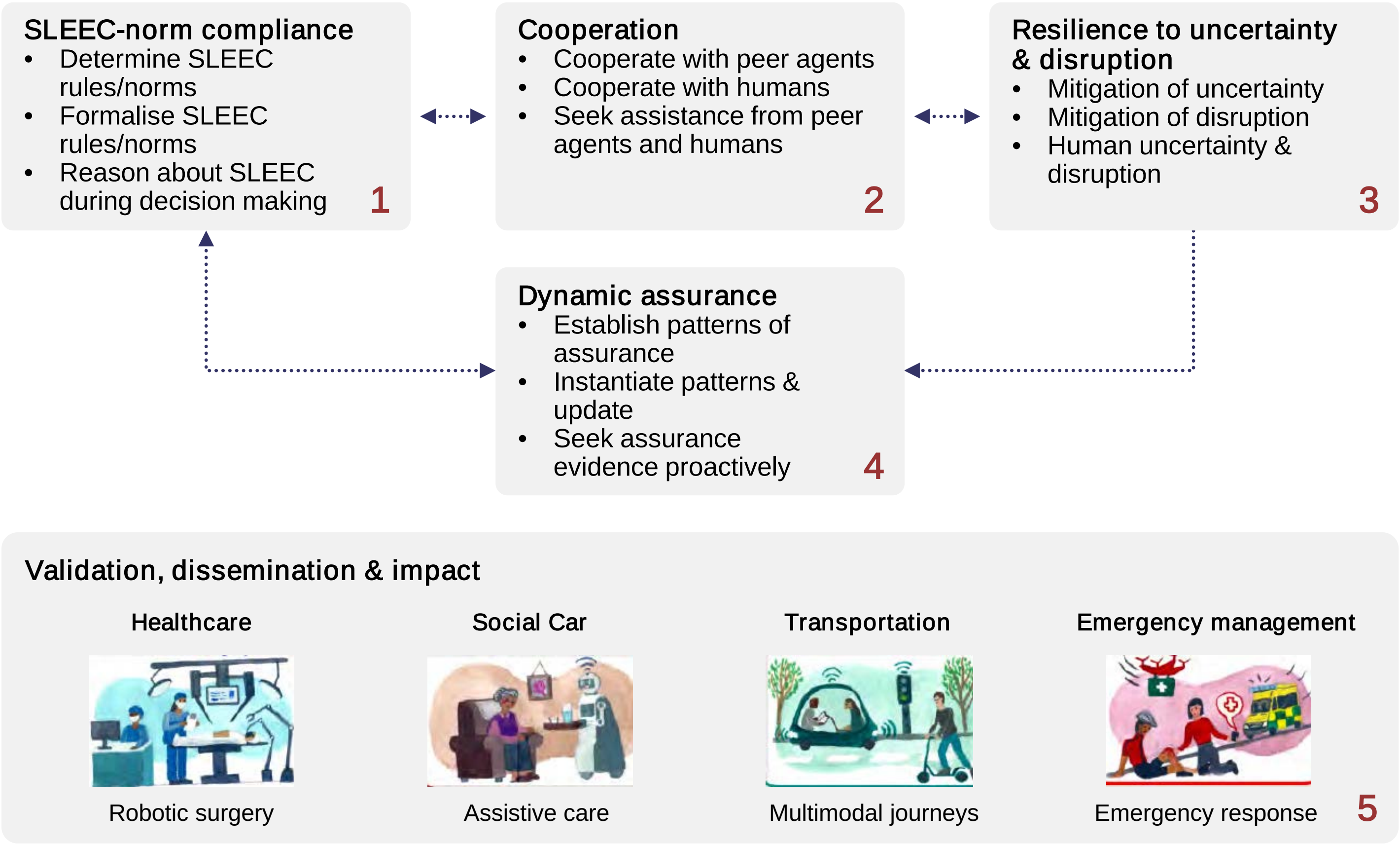
When it comes to autonomous systems, resilience is about ensuring that our machines can operate effectively, consistently and appropriately within our day-to-day lives. However, the 'real world' throws up challenges and situations that can be very difficult to predict and control. According to Professor Radu Calinescu from the University of York: "Resilience plays an essential role in ensuring that AS can be trustworthy. It is about mitigating the uncertainties and disruptions which AS encounter when deployed in real-world environments. This requires sophisticated resilience mechanisms."

The way in which the autonomous systems interact with humans and society as a whole – the socio-technical aspects – is the focus of much of the resilience work currently taking place. What are our autonomous systems required to do and what are they not allowed to do? In order for us to trust our machines, it is essential that they work in harmony with us. They must operate in ways that don't violate the norms of our society, our rules or our values. A benchmark known as SLEEC[1] - which stands for Social, Legal, Ethical, Empathetic and Cultural - encapsulates this ideology, and is based on the principles that these factors should be incorporated into the design, operation and governance of autonomous systems.

However, the reality is far more complex. Our norms and values can be subtle and nuanced. Professor Ana Cavalcanti from the University of York recounts an analogy that highlights some of these uncertainties: "If you are from the US, you know you can be arrested for trying to cross the road (jaywalking), but then if you come to the UK, you could be waiting at the roadside for a long time unnecessarily - much in the same way a robot can be brought to a halt as it doesn't understand its environment and the SLEEC norms that apply."

So how can we go about creating autonomous systems that operate within the SLEEC principles? How can we ensure that our robots always 'do the right thing'? Work is currently being carried out, both internationally and within the TAS Programme, to develop standards on the ethics of autonomous systems. We need to understand the limitations of our AS, how are they used and where. We need to understand how they behave in unpredictable environments. We also need to learn more about how they can become more resilient through collaboration with other autonomous systems.

[1] Townsend, B., Paterson, C., Arvind, T. T., Nemirovsky, G., Calinescu, R., Cavalcanti, A. L. C., Habli, I., & Thomas, A. P. (2022). From Pluralistic Normative Principles To Autonomous-agent Rules. URL: https://cutt.ly/SLEEC-rule-elicitation



SLEEC-norm compliance

- Determine SLEEC rules/norms
- Formalise SLEEC rules/norms
- Reason about SLEEC during decision making

1

Cooperation

- Cooperate with peer agents
- Cooperate with humans
- Seek assistance from peer agents and humans

2

Resilience to uncertainty & disruption

- Mitigation of uncertainty
- Mitigation of disruption
- Human uncertainty & disruption

3

Dynamic assurance

- Establish patterns of assurance
- Instantiate patterns & update
- Seek assurance evidence proactively

4

Validation, dissemination & impact

Healthcare	Social Car	Transportation	Emergency management
Robotic surgery	Assistive care	Multimodal journeys	Emergency response

5

TAS Node in Resilience Research Strands

Accounting for human input

One of the critical areas of research, however, is how autonomous systems interact with humans. They may need some of our input, a lot or none at all. They can work alongside humans or operate around us. They have the potential to operate entirely independently.

The TAS Programme is looking at how to mitigate the difficulties and unpredictable situations that arise from this. The aim is to develop autonomous systems which learn from us about how to deal with uncertainty and disruption. Humans are capable of identifying aspects of uncertainty that impact on our goals and objectives. We need our autonomous systems to do the same; to observe trends and patterns and to anticipate disruption so that they can identify potential problems, reduce uncertainty and mitigate it. This is an important area in the development of our AS – but is not at all simple, as Dr Mark Chattington from Thales UK explains:

“To build something that is trustworthy we deal with issues at the concept level which may then get magnified throughout the design cycle. So, there are certain misunderstandings or misconceptions or unknown quantities during development, and these get magnified. This is one of the biggest struggles from the industry perspective.”

Additionally, there is the element of human error. Thales have been carrying out research into the subject area of prediction and control. Amongst the findings was the discovery that an autonomous system can be working perfectly for a period of time, but then starts to exhibit undesirable behaviour due to interactions with a biased human user. The autonomous system simply re-enforces these actions, which causes the robot to fail the resilience test. The research highlighted that new tools and techniques are needed to determine resilience of the socio-technical aspects of humans and robots operating together.

Other risks need to be assessed, such as accounting for the cognitive and physical impairment of humans and the potential physical and psychological harms that can occur. How can autonomous systems communicate information to humans to avoid misunderstandings and bad outcomes?

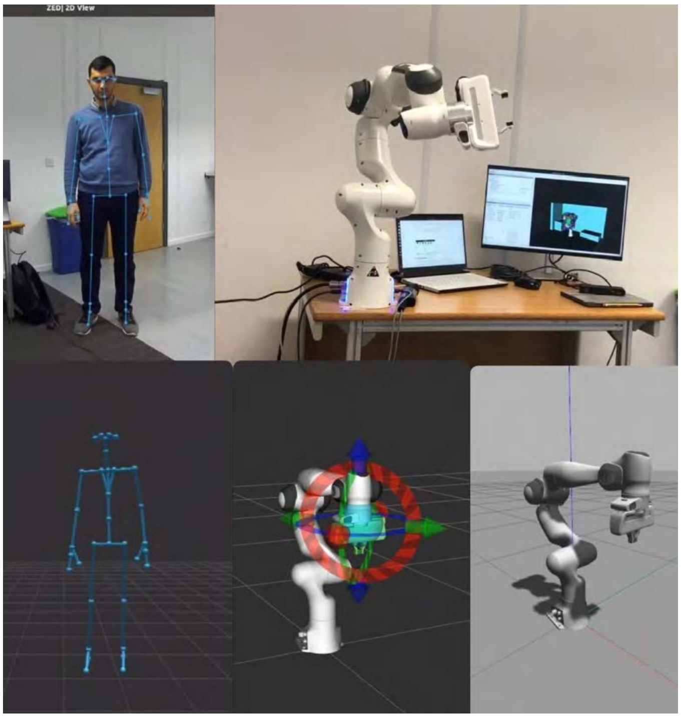
And what happens when things do go wrong? How can we be confident these systems will work and find ways to mitigate the situation as we hope? If a machine can’t be made to fix itself, the answer may be to pass the responsibility to the users, which brings up other regulatory and safety considerations.

THE IMPACTS OF RESILIENT AUTONOMOUS SYSTEMS

Research being carried out has the potential to have a real impact our day-to-day lives. An area that demonstrates this well, is the growing need for personal care. Studies are being carried out in the field of occupational therapy, observing the use of an assisted dressing machine for the mobility-impaired.

Researchers have been collecting data around how the robot interacts with patients. The ongoing study, involving seven patients and therapists, will provide information that will help shape the design of future autonomous systems.

Information being gathered includes motion data and human reactions to differing types of disruption. By understanding human intentions and analysing possible hazards, the aim is to ultimately teach the robots to ‘reason’ - understand how to predict problems and failures based on human and environmental observations. As well as for personal care, this has the potential to open up a world of opportunity for the medical and surgical fields.



University of Sheffield, Resilience case study assisted-dressing robotic arm, Franka Emika, (top right) including ROS/Gazebo simulator (bottom), with haptic feedback and AI-based motion tracking (left) (from https://www.resilience.tas.ac.uk/annual-report)

However, it’s not just the physical interactions between humans and autonomous systems that need to be taken into account. There are also should be the social and ethical complexities to consider.
In the assistive dressing example above, even basic interactions between one AS and one end-user are hard to establish and to build into rules and norms. There are further questions around the meaning of privacy or fairness, which pose important challenges to consider.

MAKING A DIFFERENCE IN THE REAL WORLD

Among the many ‘real world’ situations where resilient robots would be beneficial, several examples bring the nuances and complexities around the SLEEC[1] principles into sharp focus.

Emergency management during natural disasters, such as fires, floods, hurricanes, or earthquakes is one such case. Artificial Intelligence and autonomous systems can greatly assist human decision-making in these disaster zones and extreme environments.

However, complexities arise from there being multiple users of the robots, a range of other people involved in the relief efforts, plus unpredictable human reactions in an emergency. For example, how would a migrant at sea react to a drone (Unmanned Aerial Vehicle or UAV) above them? Or in a flood management situation, where there are different people, organisations and sources of information being used to formulate the emergency response?

Professor Mohammad Mousavi from King’s College London gives this example of the many challenges: “Machine learning may detect an object which is apparently coming towards the UAV. What is useful to do in this situation? Having awareness of what is useful to the user in various unanticipated scenarios is a big challenge.”

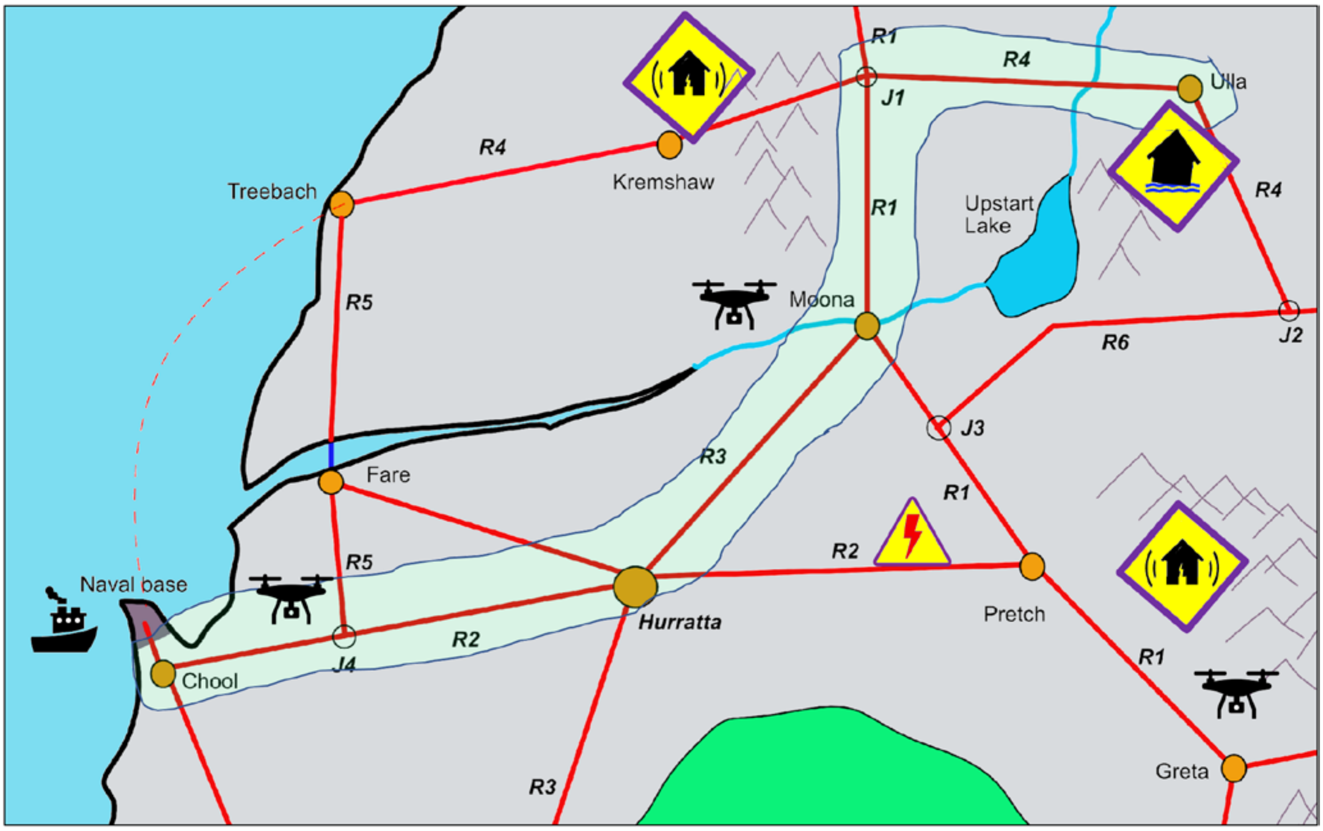
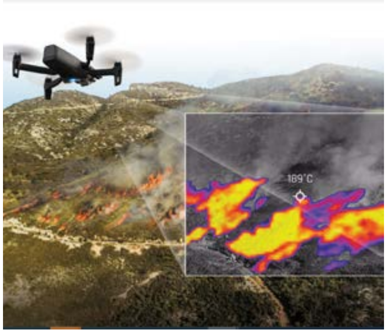


Illustration of a UAV-assisted emergency evacuation system (adapted from Paterson et al, 2019)

There is significant untapped potential for autonomous systems to be used in disaster response situations, but equally there are many variables to try to overcome.

For instance, is there a conflict between regulations and potential loss of life? This is a critical question that is currently being studied as part of ongoing research. There are collaboration issues too. It would be useful if the autonomous system identifies when it needs to delegate to a human, but ensuring ongoing cooperation between the system and the human operator is important too. To robustly consider resilience, these interactions need to be captured.



ONGOING RESEARCH AND REQUIREMENTS

Research into developing robots we can count on is continuing at a pace. Resilient autonomous systems have the potential to impact all parts of society –from space exploration to helping disabled people take part in activities that are currently inaccessible to them.

Within the TAS Programme itself, research is continuing into issues impacting resilience. Key progress has been made on the development of a SLEEC[1] Framework. Work on uncertainty, disruption, uncertainty reduction and disruption prediction has been carried out. Professor Radu Calinescu underlines its importance: "This is essential for resilience because you might not be able to mitigate all the uncertainty thrown at you and so need to obtain additional information about the environment in which you operate, which is how humans make decisions too."

Other important areas of research include communication: working with different sectors to develop a common language that is understood by all users of resilient autonomous systems. An initial version is currently being prepared for publication, but it is still early days. Professor Ana Cavalcanti explains the impact this would have: My vision is that a diverse team of people — say a sociologist, lawyer and philosopher — can use a common language as they describe the capabilities and restrictions of the AS. There will be a library of concerns and tools to identify any conflicts and resolve them. There will also be tools that the engineer can use to inform their design."

Is the future autonomous?

So, with respect to resilience, what is the ultimate goal? Can we realistically design resilient robots that complement and enhance all aspects of our lives? Will we really be able to fully rely on them when we need them? Professor Radu Calinescu believes this goal could be achieved: "My vision is of interoperable AS that self-organise into resilient, safe and beneficial 'systems of systems' capable of working alongside and for humans. For example, AS that assist with planning of evacuation routes after a natural disaster co-operating with AS that provide supplies and medical assistance to those evacuating."

However, to create truly resilient autonomous systems, we cannot rely solely on design. A key objective needs to be operational resilience. We need to ensure our autonomous systems proactively cooperate with other machines and humans, and form teams to tackle challenging problems. We also need a way to communicate the SLEEC requirements, both to developers and the robots themselves, so as they can incorporate these into their decision-making processes.

If we can achieve these goals, it could pave the way towards the creation of autonomous systems that carry out a wealth of socially-beneficial tasks in real-world environments. Professor Mohammad Mousavi adds: "It would be wonderful if we had an autonomous vehicle (AV) which could be trusted and for which we had evidence of safety and usefulness. For example, a SLEEC AV that could do the school run for me."

However, although research and development will enable us to create systems that are more and more autonomous, we need to proceed with a certain amount of caution. From society's perspective, what are we trying to achieve through automation?

Effectively, we need to weigh up the value of adding autonomy versus the risk it poses. With so much uncertainty, for example from the environment, the human user and the autonomous system itself (especially machine learning systems), we cannot assume it that is always the solution to everything; but recognise that in some environments, such as disaster management, that the risks are justified by the added value.

REFERENCES

Chance, G., Jevtic, A., Caleb-Solly, P. and Dogramadzi, S., 2017. A quantitative analysis of dressing dynamics for robotic dressing assistance. Frontiers, 4(13), p.1.

Chance, G., Camilleri, A., Winstone, B., Caleb-Solly, P. and Dogramadzi, S., 2016, June. An assistive robot to support dressing-strategies for planning and error handling. In 2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob) (pp. 774-780).

Paterson, C., Calinescu, R., Manandhar, S. and Wang, D., 2019. Using unstructured data to improve the continuous planning of critical processes involving humans. In 14th IEEE/ACM International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS) (pp. 25-31).

About the Trustworthy Autonomous Systems (TAS) Hub

The TAS Hub sits at the centre of the £33M Trustworthy Autonomous Systems Programme, funded by the UKRI Strategic Priorities Fund. Its role is to coordinate and work with six research nodes to establish a collaborative platform for the UK to enable the development of socially beneficial autonomous systems that are both trustworthy in principle and trusted in practice by individuals, society and government. For more information, please visit the website: www.tas.ac.uk

UKRI
Trustworthy
Autonomous
Systems Hub

SECURITY

KEEPING OUR AUTONOMOUS SYSTEMS SECURE IN AN UNCERTAIN WORLD

As technology advances, concerns about the safety and security of our systems increase amid ever-emerging threats. Cyber security is big business. Our personal devices require constant updating against vulnerabilities. How, therefore, do we protect our autonomous systems (AS)? How can we ensure they remain secure, especially when we are not directly involved in their operation? How do we create systems can properly assess the risks they face in different environments and respond appropriately to any issues?

For any system, security is about providing assurance that it will maintain an acceptable level of service despite any issues that might arise during operation. This is challenging enough with any technology, but with autonomous systems it is even more complex. They carry out multiple actions at once - decision-making, control, coordination and navigation - plus they operate in unpredictable environments using AI technology, which makes this is even harder.

These complex issues are among those being addressed by the UKRI Trustworthy Autonomous Systems (TAS) Programme – a £33m multi-disciplinary research programme funded as part of the Strategic Priorities Fund. Security is the focus of one of the six TAS Nodes – separate research projects examining individual aspects of trust in autonomous systems – and was the topic of one of a series of multi-disciplinary TAS workshops.

Complexities and AS Security

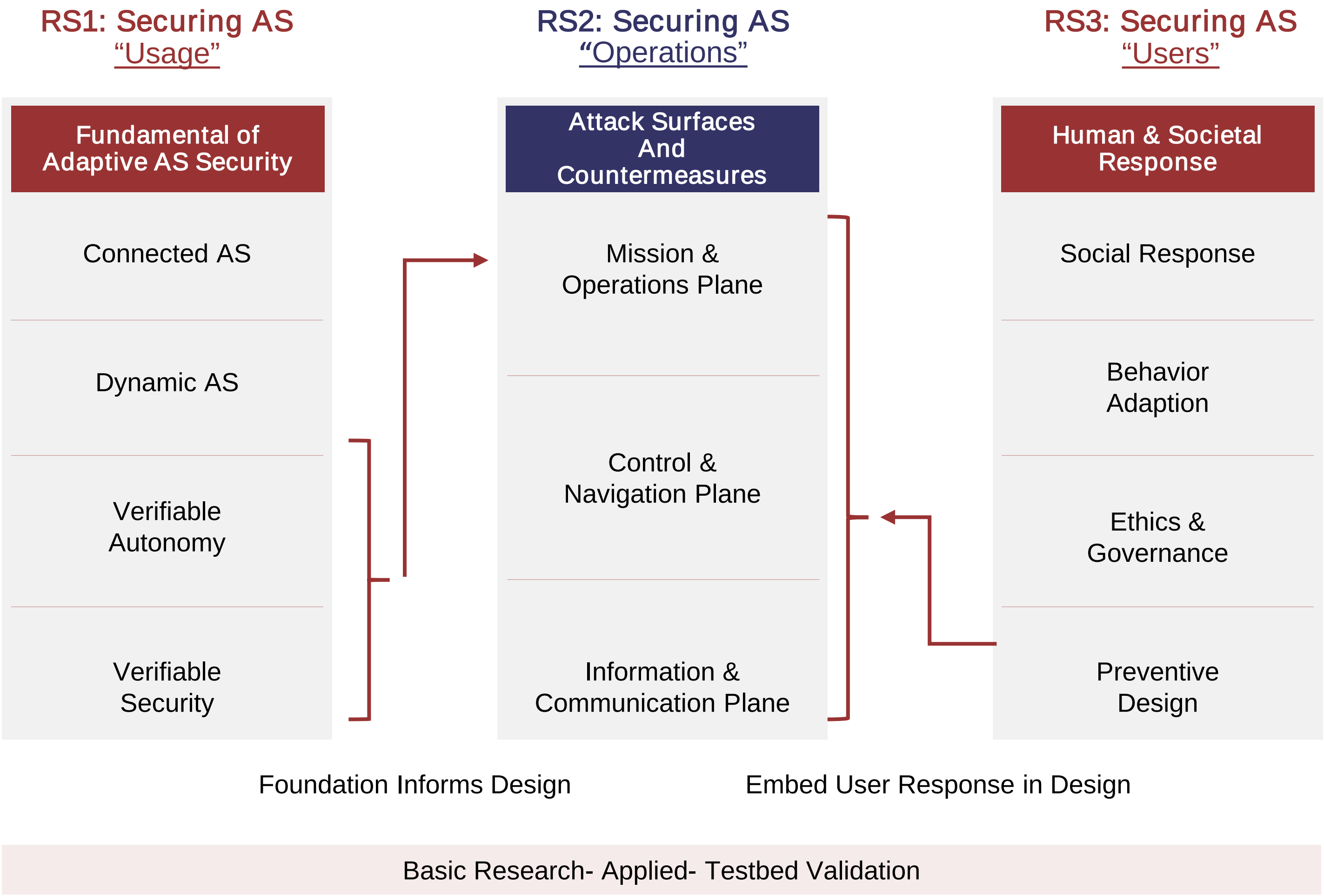
When it comes to security, what does it mean for our autonomous systems to be safe and secure? The answer is complex, nuanced and multi-dimensional - particularly when we factor in users, environmental variabilities and social impacts.

A good place to start is with specification: strict, specific parameters that determine the way a machine operates, reacts and learns. However, this is actually one of the hardest areas for TAS researchers to address, as Hamid Asgari from Thales UK explains: “Specification is the foundation of everything. Who is going to provide the specification? We need to verify the behaviour of the system based on the specification, to see whether it meets requirements or not.”

In effect, we are looking for a very structured security framework within a very unstructured environment. This is enormously challenging. Autonomous systems need to operate in a predictable manner, but they operate largely in environments where there is much uncertainty. We have to make assumptions about the threats and situations they might encounter and the behaviour they might display. We still have a great deal to learn in this area, with much of our existing knowledge being purely theoretical and based on simulations.

There are also issues with adapting the security protocols in existing technology to include autonomous systems. Many commercial organisations, for example, use information and control panels already in existence and may only be able to cope with some of the security requirements that an autonomous system demands. Professor Weisi Guo from Cranfield University says this poses real challenges: “Many commercial systems were not necessarily designed for AS. We are trying to come up with the correct requirements for autonomous systems - designing the right security protocols and new metrics which these systems will rely on.”

The TAS Programme has been examining the various security challenges in three key areas: usage, operations, users. Each area comprises ‘onion-style’ layers, which include security threats within the autonomous system itself, the AS in operation, human ‘user’ influences and the wider world. Research is underway into how threats can run across different layers and how this impacts the way that systems adapt and learn.


RS1: Securing AS "Usage"
RS2: Securing AS "Operations"
RS3: Securing AS "Users"
Fundamental of Adaptive AS Security
Connected AS
Dynamic AS
Verifiable Autonomy
Verifiable Security
Attack Surfaces And Countermeasures
Mission & Operations Plane
Control & Navigation Plane
Information & Communication Plane
Human & Societal Response
Social Response
Behavior Adaption
Ethics & Governance
Preventive Design
Foundation Informs Design
Embed User Response in Design
Basic Research- Applied- Testbed Validation

The TAS Security Node's 3 key focus areas, and the interconnected research strands (from Suri et al., 2022)

The human factor

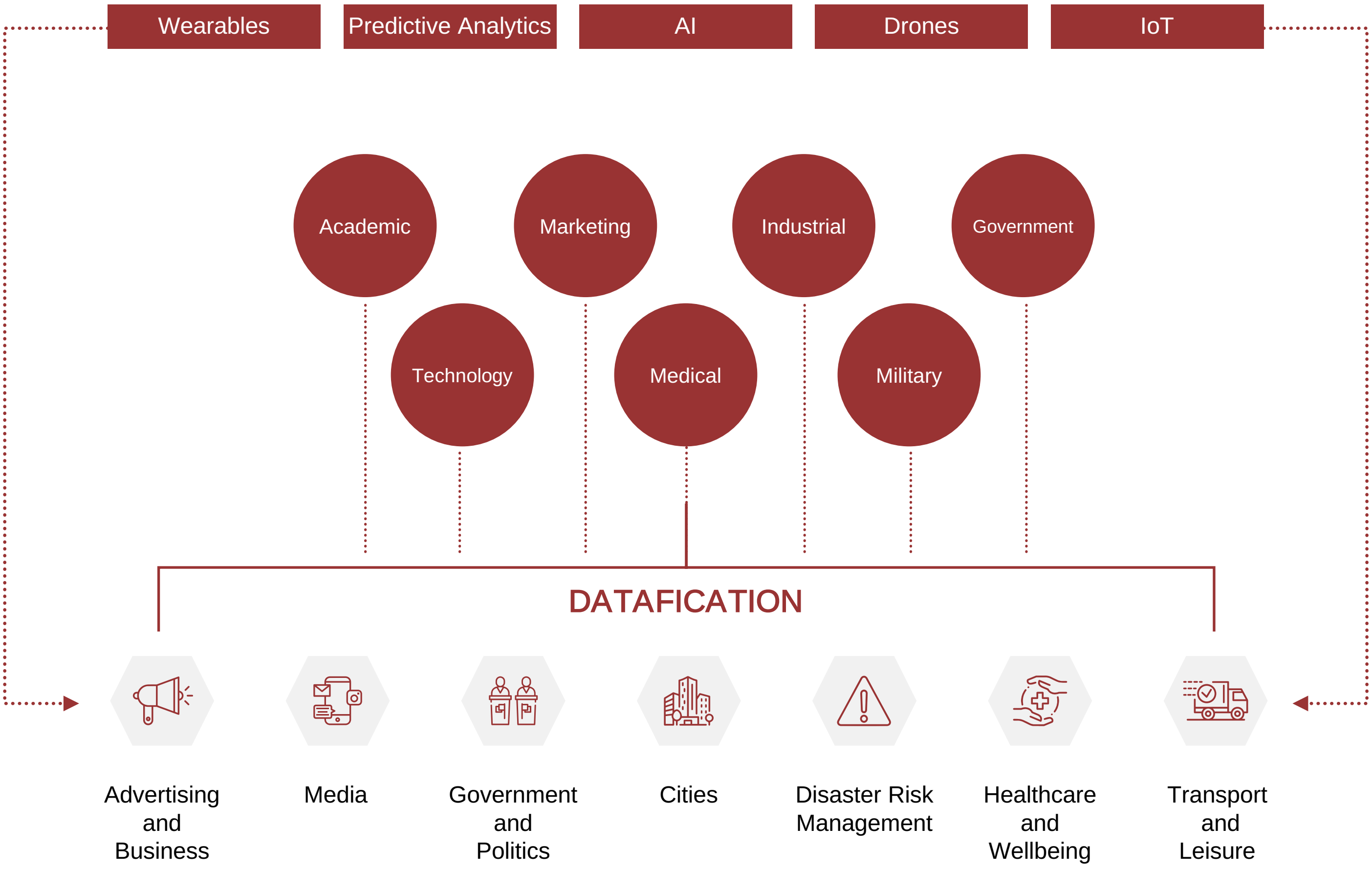
A particular area of study for the TAS Programme is examining the human and societal impacts on our autonomous systems - the social-technical aspects. How do we secure AS from new threats if they rely on human interaction? What are the ethical, legal and social implications (ELSI)?

TAS researchers are working in partnership with stakeholder organisations to co-design methods of developing awareness of the socio-technical aspects of AS security and exploring multiple ways of measuring how humans adapt to issues that arise. Co-design is central to equipping individuals, the industry and designers with tools that they can use at different points in the decision-making process. The outcomes of this work will form a toolkit for those who design, build, own and deploy autonomous systems.

Social acceptance of autonomous systems is also important to ensure wide adoption across society and industry. As users, what do we expect from our technology? Professor Jose Such from King's College London says it is important to understand what users need and think: "These AS are not in a vacuum. They will be interacting with people, and we are interested in understanding not only how people engage with, but also perceive these systems, and how you may influence this perception."

Work is being done to gather information in this area. Crowdsourcing has been used, for example, in relation to information flows and contexts within the smart home personal assistants AS ecosystem. However, information is still sparse and there are many gaps to be filled. Ethical and social aspects of AS security are commonly a secondary consideration in commercial environments, with functional and traditional security considerations often taking precedence. The challenge is how to motivate the industry to be more open and to prioritise security.

A positive example of work currently being undertaken in this field is in the design of road infrastructure. The University of Lancaster is collaborating with Highways England on an ongoing project about autonomous vehicles (AVs) and traditional road users. They are looking at how to embed ethical, legal and social considerations into a new AS co-design process. The goal is ultimately to enable driverless cars to exchange information with the infrastructure safely and securely, to improve on-board decision-making.



Cross sector e-society data science and AI/AS processing motivating the ELSI approach (from Buscher et al., 2018)

Incorporating arts and culture

The fields of arts and culture and security of autonomous systems seem, at first glance, to be worlds apart. In reality, however, there is a surprising amount of synergy, which is proving extremely helpful in the area of communication.

One of the challenges for the technology sector is ensuring a wide understanding of security issues among non-technical users. Using visuals and standardised, natural language to communicate this to a wider audience can be very effective. It is not enough to tell users what an autonomous system does and how they should interact with it. They need to be shown this in a way that is visually and linguistically understandable.

King's College London have been using film and arts to communicate with non-expert users. This example demonstrates how they used Nicolas Cage films to highlight technology in action and the threats that users face.

Film Title	keywords	Category
G-force	Hacking; Password cracking; Worm; Virus; iot attack	Hackers and cryptologists
Kick-Ass	Anonymity ; Tracing	
National Treasure	Steganography; Multiple-stage attack; Masquerading attack; Forging; Pseudonym; Social engineering; Invisible Ink; Hacking; Sensor attack; Biometrics; Integrity; Password guessing; Background Knowledge; Indistinguishability; Swapping; Tracing; Ottendorf cipher	
National Treasure; Book of Secrets	Playfair cipher, Password guessing; Device cloning; Hacking; Denial of service; Social engineering; Surveillance system; Codes	
Snowden	Cipher machines; Algorithms; Attacks; Malware; Surveillance; Trojan horse; Password Protection	
Wind talkers	Codes	
Con Air	Steganography; Authentication/Deception	Detectives and spies
Face/Off	Multi-factor authentication; Biometrics	
Gone in 60 seconds	Steganography; Codes; Insider attack	
Teen Titans GO! To the Movies	Pseudonyms; Multi-factor authentication; Biometrics; Masquerading attack; Social engineering	
Lord of war	Identity/Authentication; Confidentiality; Integrity; Social Engineering	Ordinary people
Spiderman: Into the Spider-Verse	Pseudonyms; Identity; Accountability	
The Humanity Bureau	Stolen Identity	
Ghost Rider	Biometrics	
	Anonymous	Allegory
The Sorcerer’s Apprentice	Authentication	

An example analysis (from Vigano, 2021b) of 15 Nicolas Cage films exploring the importance of the role of his films and other arts in explaining cybersecurity to non-expert audiences

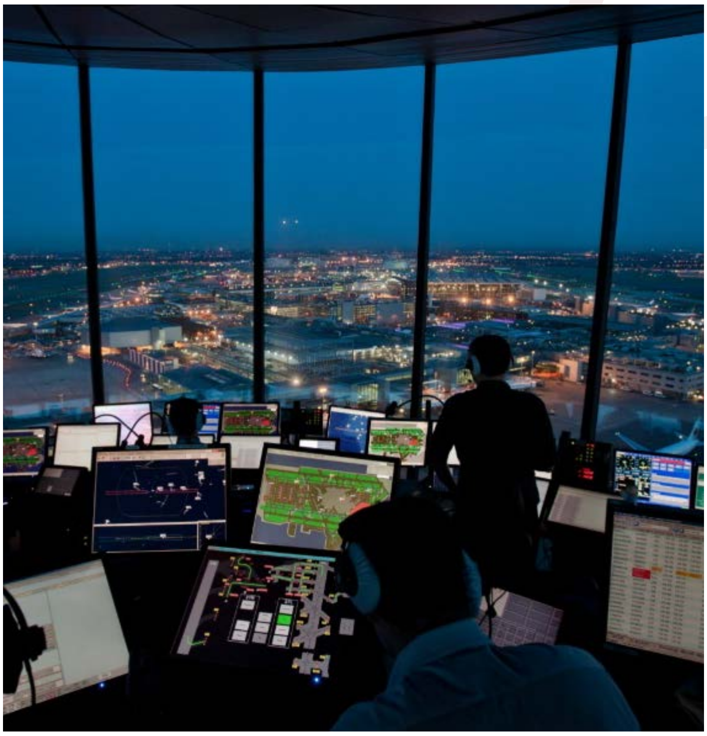
IMPACTS OF SECURITY IN THE REAL WORLD

TAS researchers have been focusing on the security of autonomous systems in two main areas: uncrewed and crewed airspace integration; and autonomous vehicles and trust among human users.

One case study they are involved with is a Heathrow Airport-led Innovate UK project, entitled Fly2Plan. The 14 partner, 15 month-long project is costing £4.6 million and aims to develop a new information-sharing model for crewed and uncrewed aerial vehicles (UAVs).

The project includes exploring secure systems using AI and other data sharing technologies for shared air and land space, air traffic management, and flight and UAV operations - including deliveries, transport, emergency response and maintenance. A secure cloud infrastructure is being adopted to replace legacy analogue systems, and more digital data and voice communication systems are being incorporated into air traffic management.

The end goal is to increase operational resilience and safety in UK airspace while reducing costs.



London Heathrow Air Traffic Control tower at night with, computer systems and human operators

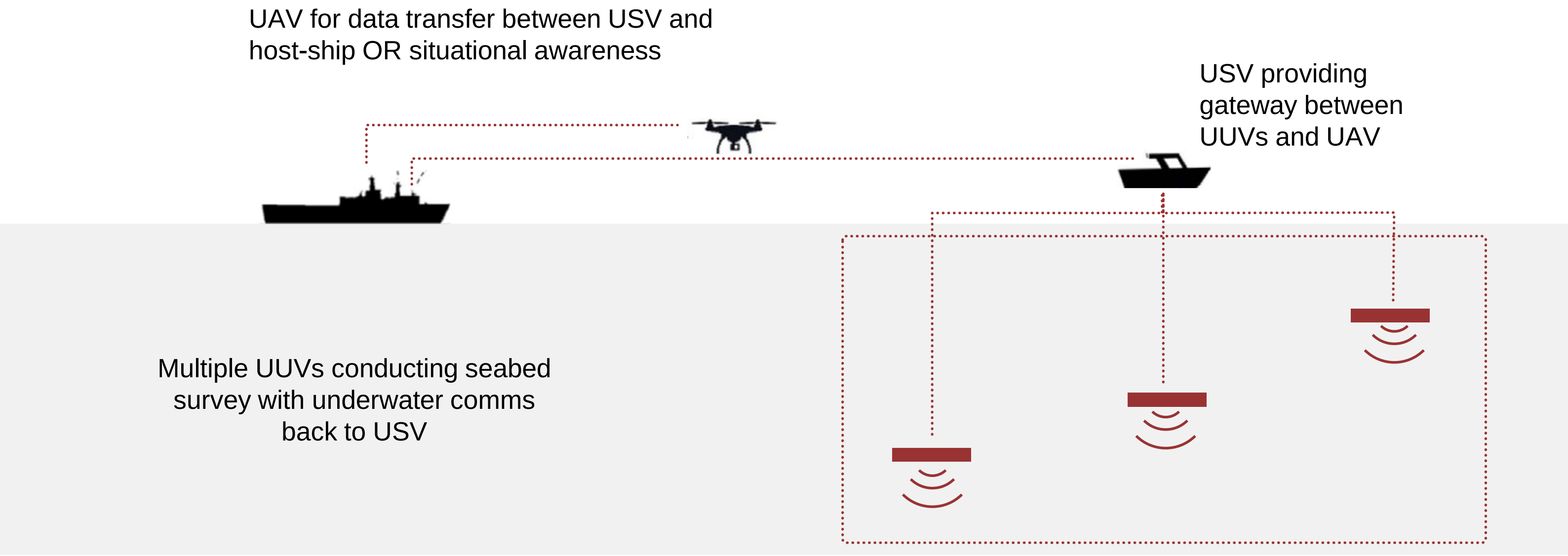
Autonomous systems working alone

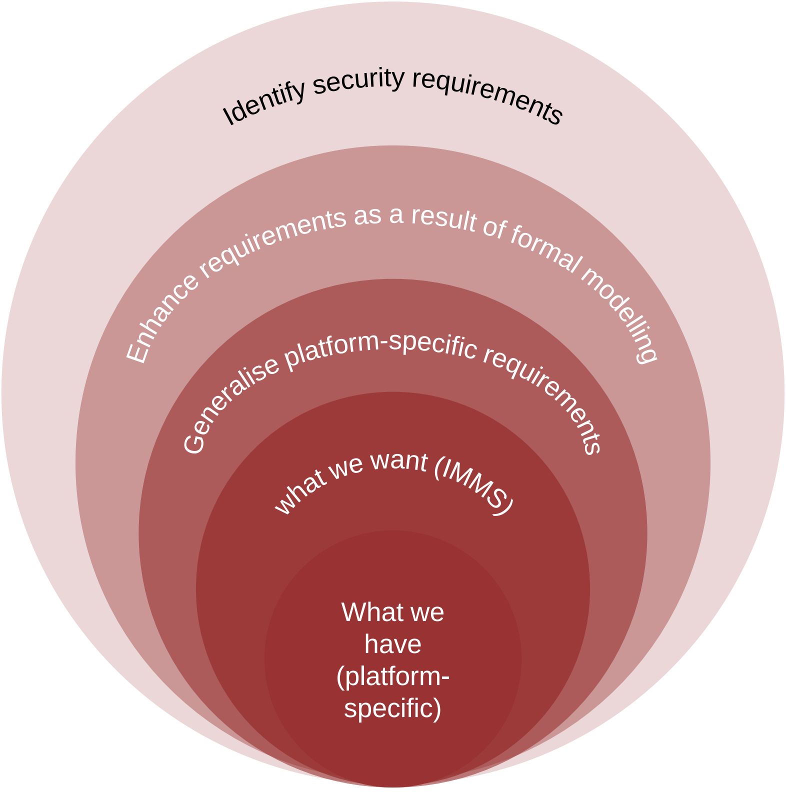
The importance of security in technology is even further amplified when humans are taken out of the loop. There is much research taking place into autonomous systems working in collaboration with their peers.

Thales are looking into various applications of multi-asset cooperative AS ‘squads’, in particular how to assess complex, real-word situations across a variety of sectors including transport, maritime, UAVs, defence and civil aviation. Researchers are looking at new approaches to identifying and analysing mission requirements in complex situations.

An example of this in marine research, where a seabed survey is taking place: an Unmanned Surface Vehicle (USV) in a shipping channel is deploying a small underwater robot (UUV), safely maintaining station and communications with a drone (UAV), before recovering the UUV.

The goal is to design robust Integrated Mission Management Systems that can supervise large squads of different autonomous systems that are working in collaboration.


UAV for data transfer between USV and host-ship OR situational awareness
USV providing gateway between UUVs and UAV
Multiple UUVs conducting seabed survey with underwater comms back to USV


Identify security requirements
Enhance requirements as a result of formal modelling
Generalise platform-specific requirements
what we want (IMMS)
What we have (platform-specific)

An AS squad use case scenario (top) and approach for eliciting requirements for the integrated mission management system (from Dghaym et al., 2021)

However, such situations do throw up a number of questions surrounding security. How do we monitor the autonomous systems during operations and how do we predict what they're doing if they go 'off- grid'? If threats or attacks occur, how do we maintain safety of both the systems and their environment?

TAS researchers have been developing models to better understand the security required, including scenarios around how critical machine-learning processes can get compromised. What happens if the learning process breaks down or the decision-making process for control, coordination, navigation and communication stops working? A robust security protocol for peer-to-peer machine learning is currently being worked on, able to tune itself to the type of data it is getting in that environment. Early results are very promising.

FUTURE THREATS AND AS DEVELOPMENT

With the challenges we face and the limited data available regarding security, how do we envisage the future? What is realistic for us? Can we really develop autonomous systems that can cope with any threat or situation, then react in the appropriate way? Do we see a future involving secure squads of autonomous vehicles that can operate without human intervention and deal with problems if they arise?

It is fair to say, we are on a journey.

According to some researchers, the future of AS security depends on the way we approach it. Professor Neeraj Suri from the University of Lancaster explains: “We need to change the paradigm. We intuitively think in terms of safety or security by design. Brittle security - wonderful if it holds, but terrible if it is compromised.”

The social, ethical and legal elements are also part of the journey. According to Professor Corinne May-Chahal from the University of Lancaster: “I think it is critical that we learn how to design-in ethics into AS. This will be critical for the development of AS. It's not just about AS- but about the functioning of AS in our everyday lives.”

There are questions too about future-proofing our systems. Can security ever keep up with the speed of innovation and usage? Professor Weisi Guo says this is extremely difficult: “Typically, you have a design life and then an operational life. This can be quite long. We don't know what will change in the next 10-20 years, so how do we design security for AS for the future?”

The journey continues, and we, the users of technology, are an important part of it. Often the push for autonomy comes from industry, but we may not be convinced that we need it or trust it. How do we reach a reasonable level of trust on these systems? We often have concerns around privacy and transparency. In order to move forward in a constructive way, it is important to inform and engage with the public about what we want and what we will accept. We need to be taken on the journey, and not simply observe as a bystander.

More engagement is needed with industry and the regulators too, to understand specific requirements and needs. Professor Luca Vigano from King’s College London explains: “Security has always been an add-on. Security by design is stated but not happening. People are realising that this is important, but we are still not seeing enterprise catching up with this.”

It is very clear that more work is needed. Technology is fast-moving; the ideas, visions and challenges for autonomous systems are multi-disciplinary and security is an ever-evolving field. What we do know for certain is that more teams and networks like the TAS Programme are much needed, in order for us to be able to create secure, integrated systems in the future.

REFERENCES

Suri, N., et al., 2022, UKRI Trustworthy Autonomous Systems Node on Security- TAS-S Annual Report 2020-2021. https://www.tas.ac.uk/News/node-in-security-annual-report/.

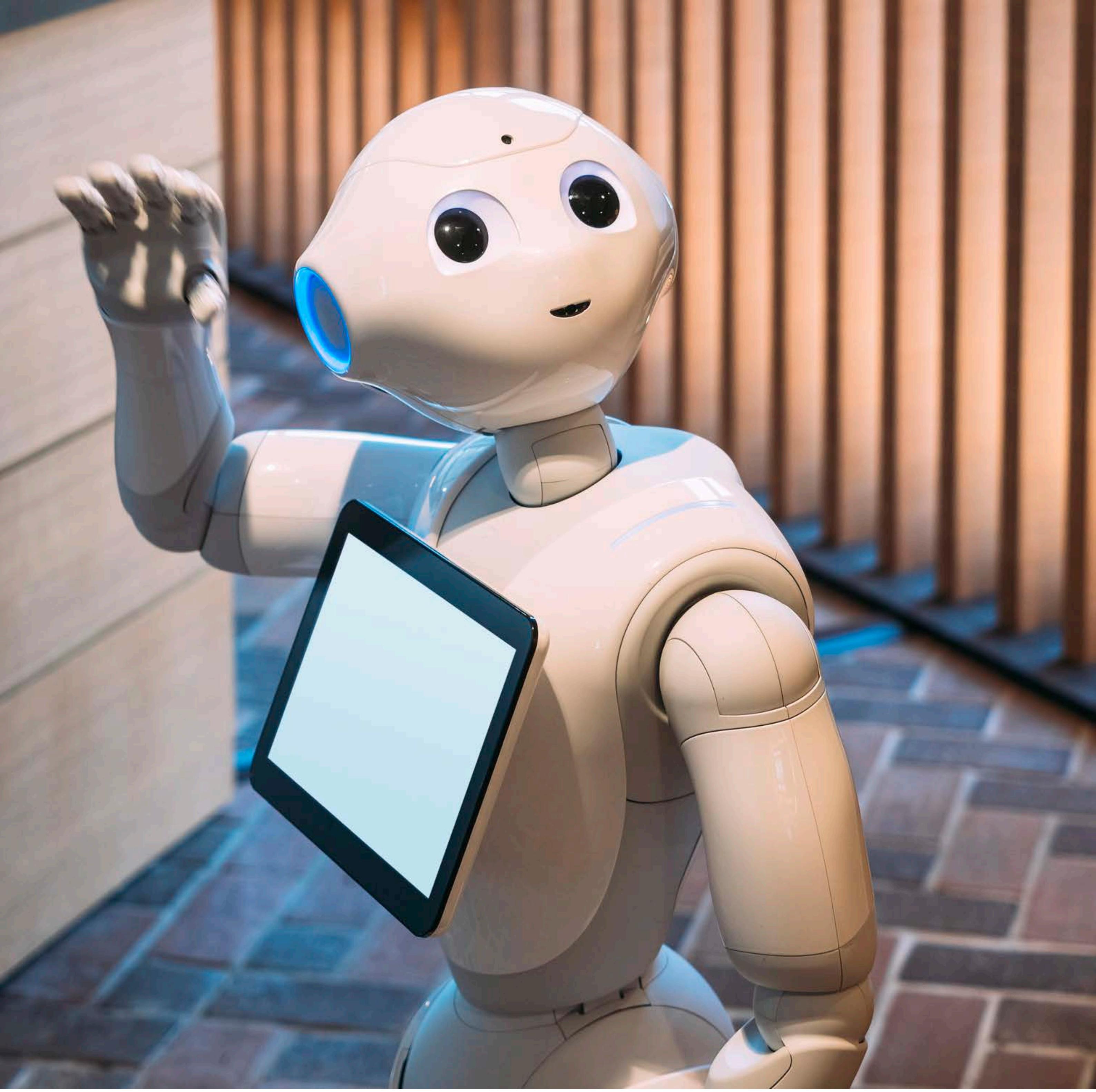
Viganò, L., 2021b. Nicolas Cage is the Center of the Cybersecurity Universe. in C Ardito, R Lanzilotti, A Malizia, A Malizia, H Petrie, A Piccinno, G Desolda & K Inkpen (eds), Human-Computer Interaction – INTERACT 2021 - 18th IFIP TC 13 International Conference, Proceedings. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12932 LNCS, Springer-Verlag Berlin Heidelberg, pp. 14-33. https://doi.org/10.1007/978-3-030-85623-6_3

Dghaym, D., Hoang, T. S., Turnock, S., Butler, M., Downes, J., Pritchard, B., 2021. An STPA-based formal composition framework for trustworthy autonomous maritime systems. Safety Science, 136(0925-7535), https://doi.org/10.1016/j.ssci.2020.105139

Buscher, M., et al., 2018. The IsITethical? Exchange: Responsible Research and Innovation for Disaster Risk Management. Proceedings of the 15th ISCRAM Conference. ed. Boersma, K.;Tomaszewski, B. Rochester: ISCRAM, 2018. pp. 254-267.
http://idl.iscram.org/files/monikabuscher/2018/2105_MonikaBuscher_etal2018.pdf

About the Trustworthy Autonomous Systems (TAS) Hub

The TAS Hub sits at the centre of the £33M Trustworthy Autonomous Systems Programme, funded by the UKRI Strategic Priorities Fund. Its role is to coordinate and work with six research nodes to establish a collaborative platform for the UK to enable the development of socially beneficial autonomous systems that are both *trustworthy in principle and trusted in practice* by individuals, society and government. For more information, please visit the website: https://www.tas.ac.uk/.



UKRI Trustworthy Autonomous Systems Hub

TRUST

Can we trust our robots? The importance of inspiring confidence in our autonomous systems

When it comes to technology, the issue of trust is critical. Without it, the future of our robots, autonomous systems and artificial intelligence is far from guaranteed. Put simply, if we don't trust our systems, we won't accept them into our everyday lives.

Trust, however, is complex and multi-faceted. It can be quickly gained and quickly lost. It is subjective – and in the context of technology, we, as humans, can be easily influenced. Therefore, evaluating trust between humans and robots is crucial and central to the work of the UKRI Trustworthy Autonomous Systems (TAS) Programme – a £33m multi-disciplinary research programme funded as part of the Strategic Priorities Fund. The Trust Node is one of six separate research projects (Nodes) looking into individual aspects of trust in autonomous systems – the overarching theme spanning all TAS projects.

The Trust Node is made up of a multidisciplinary team from the fields of AI, Robotics, Engineering, Linguistics, Psychology and Cognitive Science. Each discipline plays its own part in examining this incredibly complex issue of trust.

What is trust in autonomous systems?

In the context of technology, trust is a human attitude that an autonomous system (AS) will perform as expected and can be relied upon to reach its goal. When there is a mismatch between these expectations, trust can break down.

However, this description is far too simplistic when dealing with the nuances of human behaviour and technology. How is trust be measured and what are the factors at play? Are we able to create a common framework for trust that models a range of factors across a variety of applications and use cases?

From a technical perspective, is it possible to create a set of simple design principles to manage fluctuations in trust in autonomous systems? Can we use data to understand how trust is acquired and how it evolves with environmental factors, errors and human input? Can the behaviour of a robot be adapted if it observes an unwarranted drop in trust?

To try to address some of these questions, we need to look beyond the technology itself and adopt a multidisciplinary approach. Professor Helen Hastie from Heriot Watt University, who leads the Node on Trust, explains:

"Through multidisciplinary research, we are attempting to model phenomena observed in psychology linked to trust, developing design principles for transparent autonomous systems' interactions in order to maintain appropriate levels of trust. This is to understand the principles of how people gain trust, how it is adapted over time, to individual people and contexts, and exploring if it is possible to predict trust through both implicit and explicit signals, such as language and behaviours".

Additionally, we need to touch on a wide number of aspects of interaction and societal values. Questions such as can individual differences in approaches to trust can be measured. What is an appropriate level of trust to aim for? Can trust be predicted, and how is it influenced by factors such as context, gender and experience?

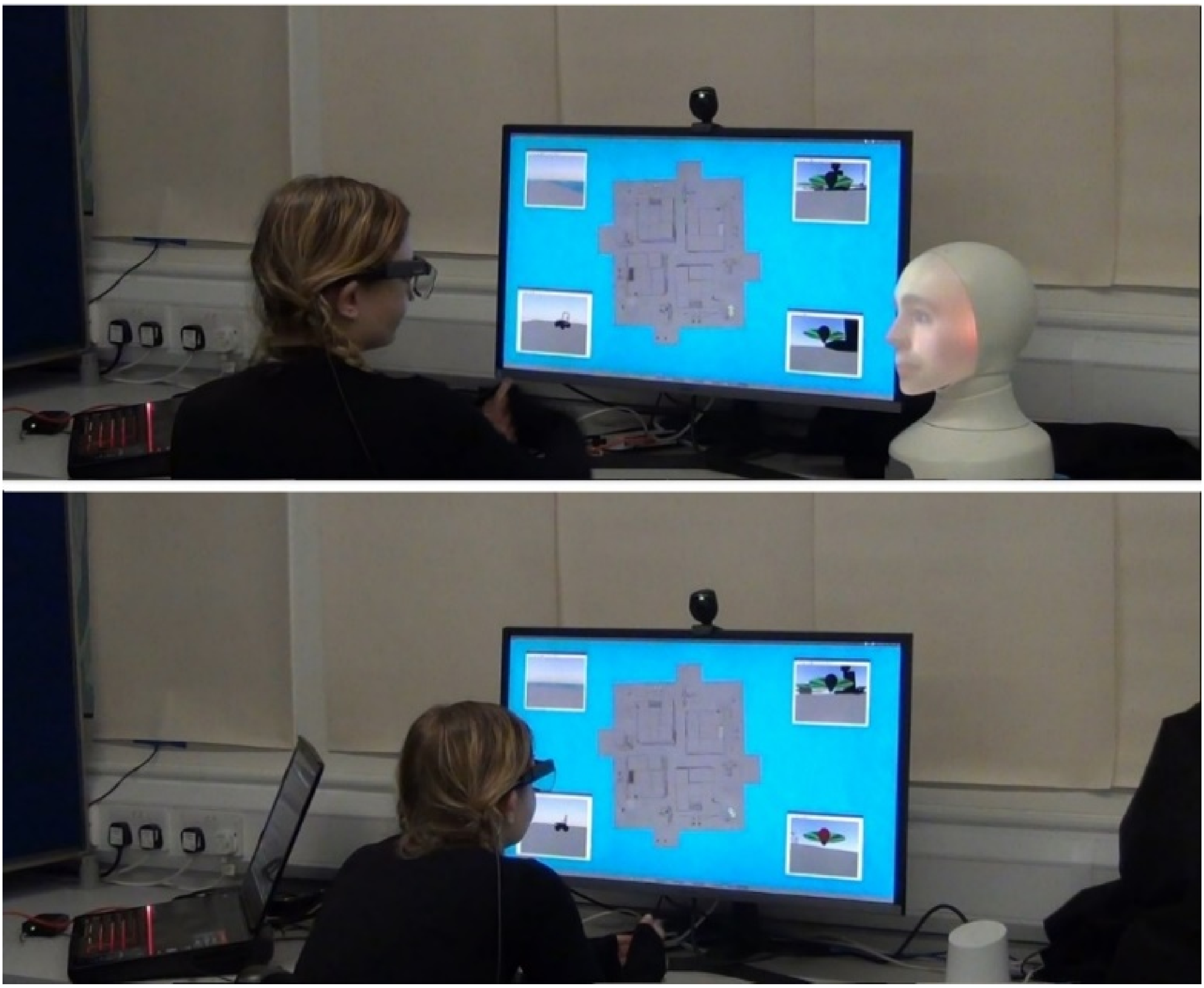
Exploring human-robot trust

The relationship between humans and robots is complex, and a multi-disciplinary approach is required to help determine how trust forms between them. TAS researchers are using a number of different methods, including modelling, automatic human trust detention and user-centred design.

Questionnaires are also proving a useful method for Human Robot Interaction (HRI) research, helping to gain insights into subjective issues such as personality traits, beliefs, predispositions and prejudices.

Much of the Trust Node’s recent HRI work has been conducted using online experiments, largely due to the limitations of the pandemic. Inspiration is taken from psychology to investigate the behaviour of users, how trust relationships are built and how the AS changes and maps to HRI changes over time.

There have been some interesting examples. One experiment contrasted a social robot with a smart speaker as a conversational agent to explore the importance of embodiment.



HRI laboratory experiments contrasting the use of the embodied robot with a smart speaker (from Robb, 2022 -paper in prep).

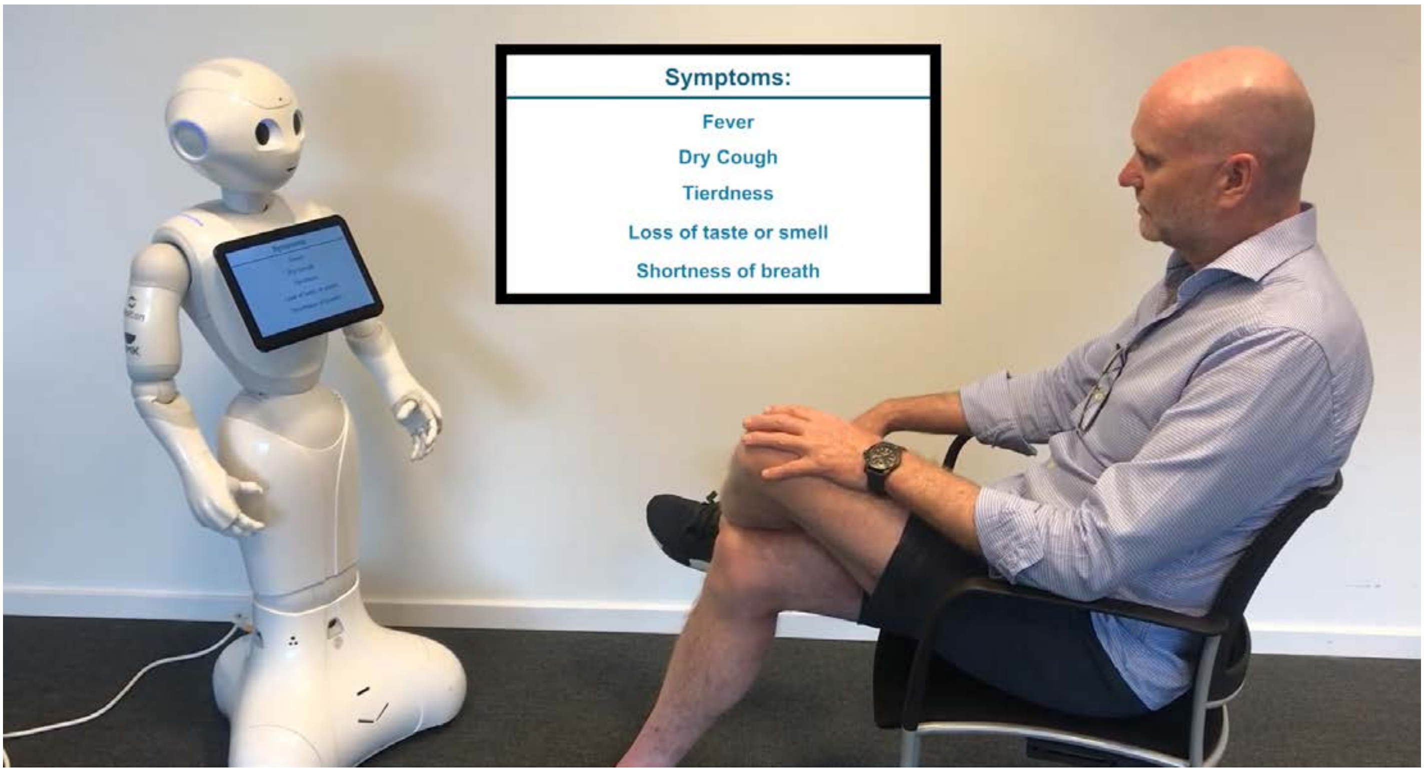
Another investigation used an online maze to explore how trust forms between humans and robots while following recommendations. A user with the goal of escaping a maze could either accept or refuse the robot's help. The study explored how our trust and behaviour was affected when interacting with three different robot personas: one neutral, one providing clear, technical explanations and one using the psychology principles of Theory of Mind (the ability to ascribe mental states to others). The research looked at how user trust changed after the robot made a deliberate mistake, how long users took to make decisions when working alongside a robot and, if they decided to follow the robot, how their decision related to their perceived trust in it (Romeo et al., 2022).

Robots in action

The pandemic also, conversely, presented opportunities for real-world research into trust. There has been a move towards using more robotic systems in various clinical settings, driven by staff capacity and the risks of close proximity working. More human-robot teams (HRT) have been introduced to support surgery and cleaning roles. This has enabled TAS researchers to gain valuable insights into trust issues towards robots, among human-robot teams including patients and carers.

A notable project centres around Pepper, a semi-humanoid robot first showcased in Japan in 2014. Today, it is a familiar ‘face’ being used in over 2000 schools and businesses spanning retail, healthcare and tourism across 70 countries.

During the pandemic, Pepper was tested as a COVID-19 triage robot, looking at people’s perspectives on AS errors, transparency and informed patient decision-making (Winfield et al. 2021). Four video scenarios of an interaction with Pepper as a triage robot were shown to an online group of crowd-sourced participants. The research looked at trust and behavioural outcomes, including how participants’ trust and decision-making changed with differing levels of designed AS transparency, both before and after a system error. The results were revealing and demonstrated that a proper understanding of the effects of AS transparency on capability is critical for ensuring appropriate levels of human trust (Nesset et al. 2021).



Video frame from vignette of an HRI interaction between an asthma patient (actor) being triaged for Covid-19 testing by an embodied Pepper robot (from https://trust.tas.ac.uk/demos).

In a different field of research, a real-world study is taking place in hospitality, with the aim of expanding the use of autonomous technology in the service industry. A Robo-Barista has been developed using a Furhat robot, which interacts using computer vision, speech and gestures. It is connected by Bluetooth to a high-end coffee machine, and a Cisco user-tiredness detection algorithm allows customisation of the strength of the coffee. Shared attention gestures and small-talk are used to engage the user. The plan is to use the Robo-Barista for long-term HRI experiments exploring how trust and attitudes towards robots and usage evolve over time.

Human user interacting with the Robo-Barista (from Lim et al., 2022)

The Furhat robot has also been used by TAS to develop an interactive robot receptionist. Featuring computer vision, it can be used to help with multiple visitor needs, such as internet access, directions, registering meeting attendees and providing general information. The Robot Receptionist has been connected to a backend visitor management system and installed in the National Robotarium in Edinburgh (Moujahid et al, 2022).

Interactive Robot Receptionist being developed for trials at the National Robotarium (from: https://trust.tas.ac.uk/demos)

Trusting in numbers

Outside of the Node on Trust, one area of interest is ‘swarm robotics’ – groups of robots working together to achieve a common goal. Swarms have huge potential to revolutionise many challenging working environments, such as emergency rescue, surveillance and logistics.

A TAS Hub-led project is underway, looking at the technical aspects of the many challenges involved in understanding and controlling groups of robots operating together in extreme environments. Working in collaboration with industry partners, and focusing on various use cases, researchers are examining key questions surrounding trust in human-swarm partnerships. How do multiple robots working together in teams make decisions – and how trustworthy are these decisions they make? Should we rely solely on their decisions? How best do we observe and monitor what they are doing? Can we use ever-advancing AI technology to increase trust and how do we evaluate this?

Building and maintaining trust in swarms is critical to maximising their potential going forward, and it is hoped that the findings will help enable their wider use across a range of different applications.

A swarm of uncrewed autonomous vehicles (UAVs)
(https://www.tas.ac.uk/part-ii-industry-driven-use-cases-for-human-swarm-interaction/)

The subjective insights into trust

So where do these TAS research projects lead us on the question of trusting our robots?

Early results are providing some valuable insights into the challenges surrounding human subjectivity, psychology and attitudes.

Research carried out using the Propensity to Trust (PPT) questionnaires showed that our personality traits do indeed influence whether we trust an autonomous system and if we perceive it to be trustworthy. The next step is to understand how to use these results; for example, looking at how to adapt the robot’s interactions based on users with different PTT levels.

Context is also a very important factor. Personality-matching theories do not always apply in every situation and the human-robot relationship is greatly influenced by situation, environment and expectation. In the case of the Robo-Barista, for example, all the participants preferred the extrovert robot to the introvert one. This could be down to stereotyping or pre-conditioning – after all, getting coffee is a social activity and we expect an engaging interaction.

Robots with clear, technical explanations – known as transparent robots – are generally more trusted. When following directions in the HRI Maze study, people trusted the robots that gave more direct responses, and these users were also likely to make better decisions. This is a key finding in the understanding of human-robot collaboration and AS adoption going forward: clear communication inspires trust.

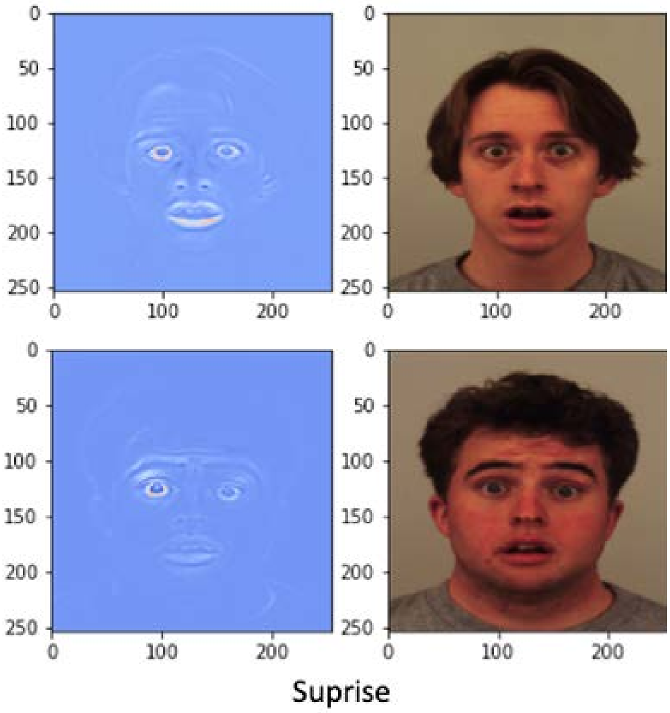
However, there is a balance to be struck and too much communication has the opposite effect. If a robot constantly requires too much feedback or asks too many questions, we tend to lose trust in it.

Trevor Woolven from Thales UK explains the theory: "Fundamentally with autonomy we are trying to do more, with less equipment and fewer humans – and faster. If humans have a machine that is taking up time and slowing things down, then they are going to turn it off."

Why feelings matter

While it is clear that psychology really does influence our decision-making and trust, emotions also play their part – and steps are already being taken to enable interpretation in this field.

A project is underway into improving robot understanding of human emotions using biometrics, vision systems and indicators for language. The TAS Trust Node has developed Deep Neural Network models for facial emotion recognition that, unusually, provide a degree of explainability and transparency on AS prediction. The project involves the extensive study of facial expressions. Heatmaps are generated of different emotions and presented back to the user on the screen of a Pepper robot. It is hoped the research will help increase automated trust detection, prediction and robot adaptation. However, as with all facial recognition technologies, this also raises important questions around privacy protection, and the risk of biases in the training data leading to misclassifications.



An example heatmap for explainable facial emotion (Surprise) recognition, presented back to the human participant to build Trust (from Zhu et al., 2022).

Can we trust the future?

What does the future hold for our autonomous systems? Will we ever place our full trust in robots and welcome new advancements and technology with open arms?

There is still much to learn: how we teach our robots, the common language we use, what happens to our trust if the robot makes a mistake, how we restore trust, how robots can work with us. According to Trevor Woolven, ongoing studies are essential to the wider expansion of AS into society:

"Research into trust is fundamental to the deployment of autonomous robotic systems in the military, air traffic control and autonomous vehicles spheres. The sensing and moving problems can probably be solved, but if we don't address the issue of trust they will never be used."

So, what will enable us to fully trust autonomous systems? It is the question loaded with complexities that spans the entire TAS Programme. Will full trust finally be gained when we make our robots react and respond as we would? Humans are subjective and capable of making errors or the wrong decisions. Would we forgive our technology for displaying the same qualities? Do our robots need to be more reliable than we are ourselves?

To what extent do we need to see our robots as extensions of ourselves, to have personalities? Can we do too good a job of creating a thinking, intelligent robot? In effect, can a robot be perceived to be too capable?

Dr Mei Yii Lim from Heriot Watt University thinks it is a possibility that needs to be explored: “Interesting questions are raised about AI versus trust. We are making these machines more intelligent and smarter, even able to model the user’s trust. At some point we may achieve what we are aiming to - but is there a tipping point where the user starts to distrust the AI as a result as they think it is going to outsmart them? I believe there is – and it is important we find out about this.”

At present, extensive research continues into trying to address some of these important questions – and exciting projects showing the positive impacts of autonomous systems in our lives are helping build trust among wider society. We are looking into a future where our robots not only help us manage dangerous situations, assist our surgeons and support our healthcare systems, but also ‘live and work’ in close proximity to us, helping us with tasks such as dressing, feeding and companionship.

For this, our autonomous systems need to instil trust – and we need to understand how to make that happen. It is fundamental to their adoption into our lives. We can make technological advances, but we need to be comfortable and fully trust them to really help shape our future.

References:

Robb. D., 2022. Slides prepared for the ‘TAS Programme Trust Workshop’. 31st January 2022.

Romeo,. M**,** et al., in prep. Exploring Theory of Mind for Human-Robot Collaboration. In submission to 31st IEEE RO-MAN 2022. International Conference on Robot & Human Interactive Communication. Naples August 2022.

Winfield, A.F.T., Booth, S., Dennis, L.A., Egawa, T., Hastie, H., Jacobs, N., Muttram, R., Olszewska, J. I., Rajabiyazdi, F., Theodorou, A., Underwood, M., Wortham, R.H., and Watson, E.(2021). IEEE P7001: a proposed standard on Transparency. In Frontiers in Robotics and AI, section Ethics in Robotics and Artificial Intelligence

Nesset, B., Robb, D.A., Lopes, J., Hastie, H., 2021. Transparency in HRI: Trust and Decision Making in the Face of Robot Errors. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI’21). HRI '21 Companion: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. March 2021 Pages 313–317 https://doi.org/10.1145/3434074.3447183

Lim, M.Y., Lopes, J.D.A., Robb, D.A., Wilson, B.W., Moujahid, M., Hastie, H., 2022. Demonstration of a Robo-Barista for In the Wild Interaction. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI’22). https://dl.acm.org/doi/abs/10.5555/3523760.3523974

Moujahid, M., Hastie, H., Lemon, O., 2022. Multi-party interaction with a Robot Receptionist. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI’22) https://dl.acm.org/doi/10.5555/3523760.3523907

Zhu, H., Yu, C., Cangelosi, A., 2022. Explainable Emotion Recognition for Trustworthy Human-Robot Interaction. Context-Awareness in Human-Robot Interaction Approaches and Challenges Workshop on IEEE/ACM HRI 2022. https://www.researchgate.net/publication/359067541_Explainable_Emotion_Recognition_for_Trustworthy_Human-Robot_Interaction

UKRI Trustworthy Autonomous Systems Hub

VERIFIABILITY

CHECKS, BALANCES AND GUARANTEES – HOW VERIFICATION IS CRITICAL TO THE CREATION OF TRUSTWORTHY TECHNOLOGY

Autonomous systems are predicted to play a central role in our day-to-day lives in the future, with technology such as drones, driverless cars and assistive care robotics already making a huge impact. It is an exciting prospect, but not without its concerns. Without humans to operate them, how can we be sure that our autonomous systems are going to function in unconstrained and complex environments as we need them to? To fully welcome them into our lives, we need to be confident in the safety and reliability of their decision-making in the presence of uncertainty and while comprising artificial intelligence components. We need assurances.

Determining trust is central to the work of the Trustworthy Autonomous Systems (TAS) Programme – a £33m multi-disciplinary research programme funded as part of the UKRI Strategic Priorities Fund. The UKRI TAS Programme comprises six Nodes, which are separate research projects, each examining individual aspects of trust in autonomous systems. Among these is the TAS Verifiability Node, which is exploring the tools and techniques we can use to assess our autonomous systems and give us confidence in their abilities.

There are many challenges to be overcome in this context. Is it really possible to guarantee that our smart technology is consistently reliable, safe and secure? What is the best approach to verification? What checks and balances can we put in place to ensure our autonomous systems meet expectations? How best can we guarantee their operation and decision-making throughout their life cycle?

As with many areas of autonomous systems (AS), artificial intelligence (AI) and machine learning (ML), verification is not straightforward.

The complexities of assurance

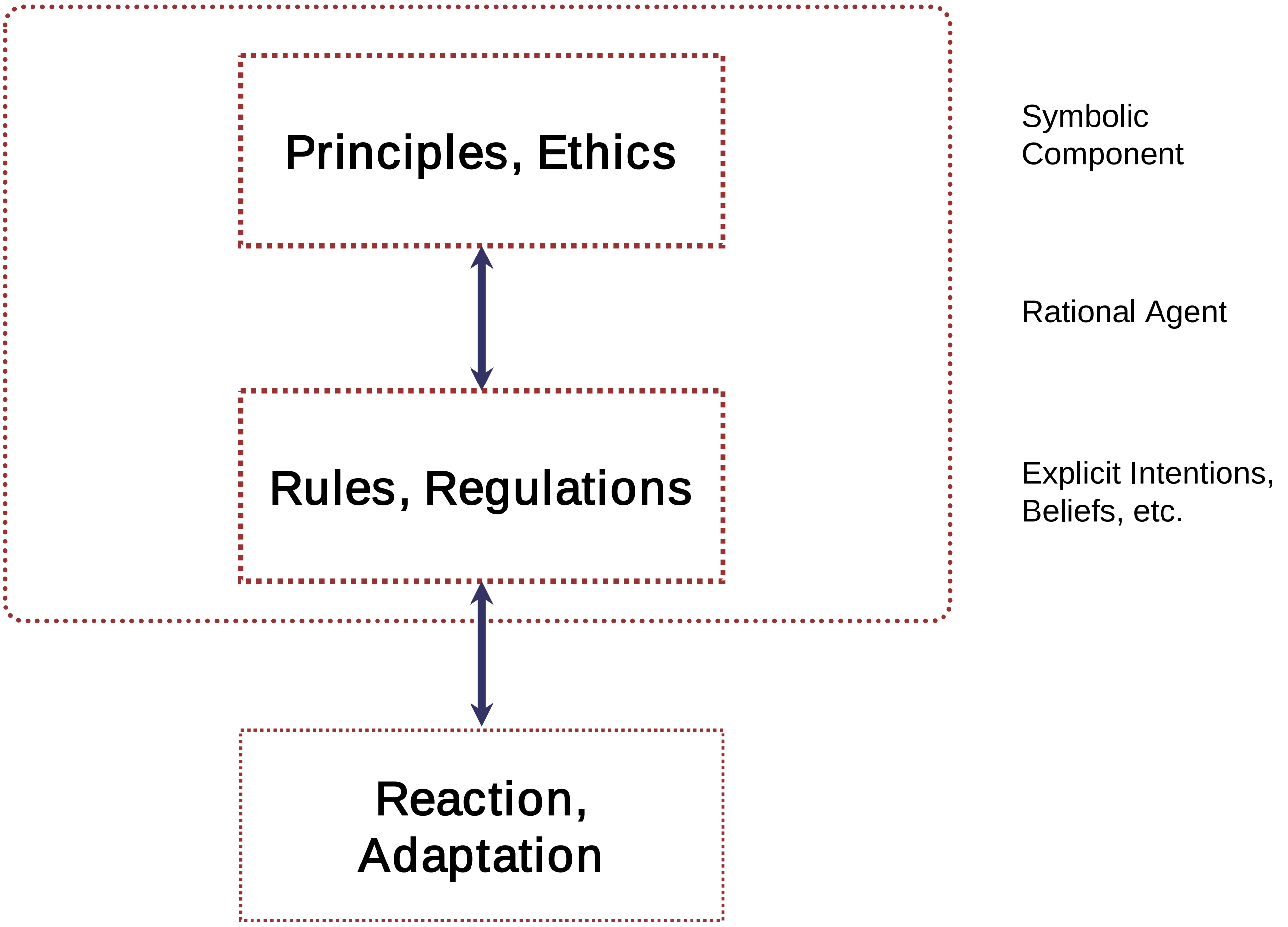
One of the main challenges of verification is what's known as heterogeneity; the fact that autonomous systems all have differing software, hardware and application environments. A one-size-fits-all approach does not work. Verifying AS involves many different areas of expertise. What do we verify in terms of a system's properties? How can we ensure these systems are fault-tolerant and adaptive? How can we efficiently upgrade them? And how do we interpret results and know if we have succeeded?

A big challenge is making the step from cyber physical systems to autonomous systems. How can we apply verification techniques used in traditional computer science to AI-based systems? Software is now making decisions that humans used to make. We need to be confident that we can trust these decisions, especially when they are mission-critical and in unpredictable environments. We need to assure safety. How do we do this?

According to Hamid Asgari from Thales UK, who are involved in verification research, the best approach is a holistic one: "Incorporation of AI into AS will need new processes and techniques from the start, across the whole lifecycle from context, through design, development, testing, integration, runtime evaluation and verification."

There is also the human factor. How much input - if any - will we have in any scenario, and what impact will this have? Ethical concerns also need to be taken into account during the verification process, which poses interesting challenges. Many real-life situations are nuanced, and we make decisions based on circumstances, multiple sources of information and social conventions. Would an autonomous system act in the same way as a human in a complex situation - and how can this be verified? As an example, would an autonomous vehicle waiting at a road junction make the same decision we would? Would it follow the highway code at all costs? How would we verify its decision?

In the face of these complexities, it becomes clear that we need to look at verification in a broad context across the entire AS hierarchy. In this way, we can try to understand not just what decision was made, but why it was made.



The Hierarchy of Autonomous System decision making from low and mid-level autonomy to high level human in the loop decisions (extract from white paper (Fisher, 2021).

VERIFICATION TECHNIQUES

The various types of software, hardware and uses for our technology means that, in reality, there is no single technique to verify any one system.

Different approaches to verification are also required throughout a system's lifecycle, from the design stage right through to the online verification of the deployed system. Once in operation, there is a need to verify the decisions that autonomous systems make in uncertain and changing environments. In addition, the verification technique used depends on the assurances required. For example, if an absolute guarantee for a critical system is required, mathematical proof may be needed. For other systems, a statistical guarantee may be sufficient for the specification.

Model-based testing is currently the most widely-used technique for verification, but due to the heterogeneous nature of AS, finding models from different disciplines to enable verification is a big challenge. How can we build and specify verification test cases, create robust simulations and representative synthetic environments? There are challenges such as requiring good specification in the first place - something that's even harder for AI systems. Formal modelling expertise and the ability to abstract away from unimportant details.

Dr Son Hoang from the University of Southampton believes the answer lies in controlled simplification: “Model Abstraction is key for successful verification. If the model is too abstract you cannot verify properly, if too concrete it would be impossible to verify practically.”

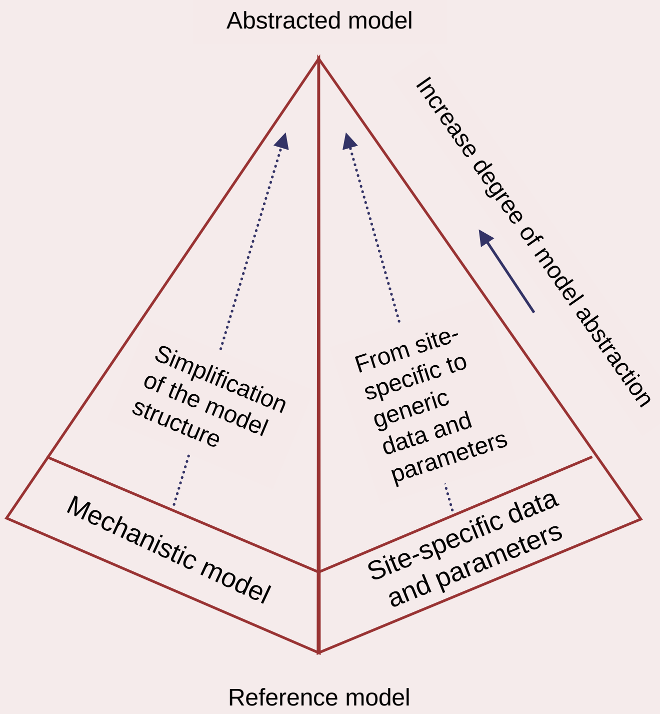

Abstracted model
Increase degree of model abstraction
Simplification of the model structure
From site-specific to generic data and parameters
Mechanistic model
Site-specific data and parameters
Reference model

Diagram illustrating model abstraction. The base of the pyramid represents the most detailed model, often the reference model, as you move up the pyramid the model becomes more abstracted both in terms of the model structure and dataset (Schneider et al., 2010)

Verification in action

It is fair to say that verification is a continuous process that is constantly evolving. However, encouraging progress is being made.

Researchers at the TAS Verifiability Node are currently collaborating with other TAS Nodes and industry partners to gather data which will influence future tools and techniques.

Academic research projects are underway that are proving to be important case studies for verifiability. One such study focuses on a fire-fighting drone. This project at the University of Leeds, is looking at the use of trustworthy autonomous systems in emergency situations. The focus is on the use of AI to help tackle fires in high-rise buildings using a simple computer vision-based ‘search, position hold and extinguish’ implementation method. TAS researchers have been working with the project teams on modelling, with simulations and flight experiments expected in the future.

University of Leeds Fire-fighting drone Verifiability Case Study (from https://bit.ly/Verifiability)

This work will help pave the way for other potential applications in an emergency response situation; for example, supporting refugees within camps or search-and-rescue operations at sea. These are unpredictable scenarios where verification is needed to provide assurances and affirm the safety of the system. Professor Radu Calinescu from the University of York explains:
“If the presence of the robot is unknown to the human being supported, the interaction cannot be planned in advance, so ensuring safety, security and compliance with legal and ethical norms all need to be verified.”

Potential uses of verified autonomous systems are not, however, limited to the emergency services. Another field with great potential is healthcare.

TAS researchers across multiple Nodes have been working on a case study with the University of Sheffield involving an assistive dressing robot. Using AI, the robot would help a patient put on a jacket using a single or bi-manual robotic system with sensors, touch-based and speech-based feedback, computer vision and prediction of the patient's position and movement. This is a significant case study for verification research, as it involves the use of real robots which the TAS researchers can model. Work is underway to provide safety assurances using a range of tools and techniques. Modelling of human and environmental interactions and robotic behaviour is also taking place.

However, despite having real-life examples to work with, verification is by no means straightforward.

Robotic-assisted care for people with physical impairments needs to factor in complex interactions between the robot and the patient. For domestic, social and healthcare robots, the issues go beyond safety and security to include privacy, ethics, transparency, explainability and regulatory barriers. This case study perfectly illustrates the importance of verification for wider acceptance of technology in society. It is hoped the findings will provide vital information on these aspects and help 'stress test' the decisions that the robots need to make.

University of Sheffield assisted-care robotic arm Verifiability case study (from https://bit.ly/Verifiability)

Next-level robot teams

Another exciting area that will push the capabilities of verification is swarm robotics – multiple robots acting together as a system. With swarms, there are complexities at all levels; from the design, deployment and formal verification of the properties, to the behaviour of the robots working together. Re-creating true-to-life outdoor scenarios in a simulated environment is challenging. There are regulatory standards for individual robots, but none currently exist for robotic swarms.

It is still early days for this emerging technology, but various studies are already underway.

Thales are working on a project to verify the behaviour of a fleet of drones, including verifying some of the hardware that will be running the operating system and software. They are also looking at verifying the interaction between the drones and the actions of the fleet. This is not something that can be done in the software alone, as one drone acting alone cannot fulfil the mission; it depends on the interaction of all the drones. Formally verifying this kind of scenario is challenging.

CAN WE GUARANTEE THE FUTURE?

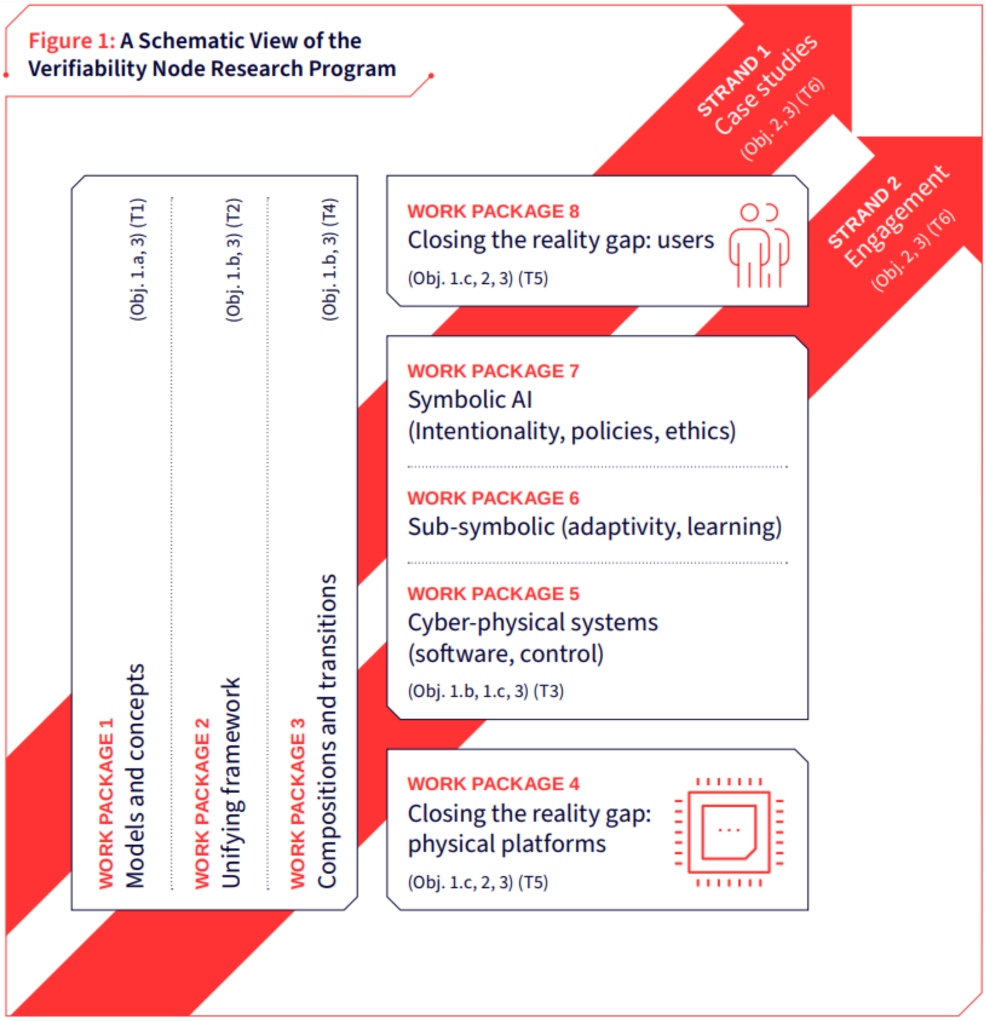
In recent years we have all begun to accept more and more intelligent technology into our everyday lives, from helpful to life-dependant: smart speakers, chatbots, delivery drones and driverless trains to robot-assisted surgeries.

But as advances in AI and AS gather pace, verification of these emerging systems will be crucial to their integration into society. Trusted testing and reliability is key to this - but achieving it is complex. Significant strides are, however, being made and collaboration is at the heart of this.

The TAS Verifiability Node has been casting its net wide, reaching out to AS researchers in the UK and internationally, building a community of people involved in verification to share expertise and information. They have also been taking part in government consultations and engaging with the wider public to expand knowledge.

As for verifiability itself, the vision is a holistic one – a unifying framework where domain experts from different disciplines, from human aspects to roboticists, can interact with the same verification framework using their different domain-specific languages and notations. It needs to be a framework that is understandable for all and accessible for any type of industry.

According to Professor Mohammad Mousavi from King’s College London it is about creating “one tool that connects all these different aspects and that comes up with the holistic verification result – not only for the initial design but also throughout the lifetime of the system. This common verification framework will provide continuous trustworthiness guarantees.”



A schematic of the Verifiability Node Research Program (from https://bit.ly/Verifiability)

This highly-integrated framework would also address another key issue – that of the cost. It would help reduce the financial, computational and energy implications of verification, even with the increase in complexity that is likely to be required in the coming ten years.

According to Professor Jim Woodcock of the University of York, the impact would be significant: “The cost of verification will decrease 100-fold for the same level of trustworthiness, or better, and it will be scalable. This depends on understanding problems, notations, proof techniques, mechanisation technology, model checking but also theorem proving.”

As for the future and our acceptance of ever-more autonomous systems, researchers believe we need to look closely at what we really want to achieve. Professor Kerstin Eder from the University of Bristol explains: “Rather than push limits of machine learning and autonomous systems, we need to reassess the way we design and engineer these systems. For safety-critical systems we need to fundamentally understand what we can or cannot do and be clear about that.”

In effect, it is about accepting trade-offs that will create trust: using safety-critical technology in situations when it is required and pushing the limits of AS, AI and ML in cases where it is not.

If advances in technology depend on public confidence, it critical that we build that trust through assurances so its future can be guaranteed.

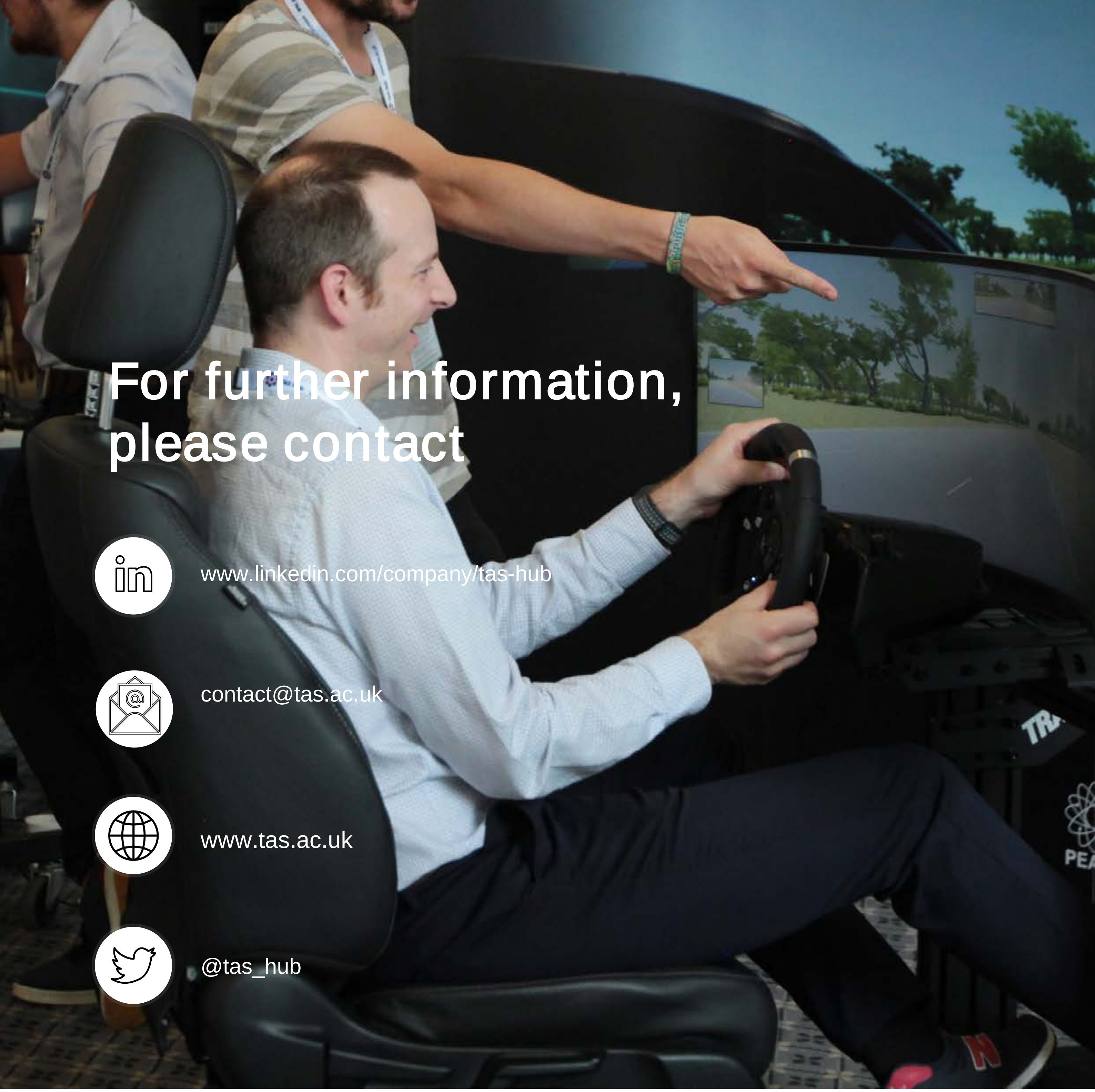
Our Definitions

Autonomous System: A system involving software applications, machines, and people, that is able to take actions with little or no human supervision.

Trust in Autonomous Systems: Trust is defined in many ways by different research disciplines. The TAS programme focuses on those notions that concern the relationship between humans (individuals and organisations) and autonomous systems.

Trustworthy Autonomous Systems are trustworthy when their design, engineering, and operation ensures they generate positive outcomes and mitigate potentially harmful outcomes. Whether they are trusted depends on a number of factors including but not limited to:

- Their robustness in dynamic and uncertain environments.
- The assurance of their design and operation through verification and validation processes.
- The confidence they inspire as they evolve their functionality.
- Their explainability, accountability, and understandability to a diverse set of users.
- Their defences against attacks on the systems, users, and the environment they are deployed in. Their governance and the regulation of their design and operation.
- The consideration of human values and ethics in their development and use.



For further information, please contact

www.linkedin.com/company/tas-hub

contact@tas.ac.uk

www.tas.ac.uk

@tas_hub

UKRI
Trustworthy
Autonomous
Systems Hub

UK Research
and Innovation