Trusting Machines? Cross-sector lessons from Healthcare & Security

30 Jun 2021 - 2 Jul 2021

Conference report

Mohammad Naiseh[1], Jediah Clark[1], Sylvaine Tuncer[2],

Mohammad Divband Soorati[1], David Bossens[1]

[1] University of Southampton, Southampton, United Kingdom

[2] King's College London, London, United Kingdom

Contents

Conference Scope

RUSI and **UKRI TAS Hub** – Trustworthy Autonomous Systems Hub – have presented ***Trusting Machines? Cross-sector Lessons from Healthcare and Security***. The conference was held between June 30th and July 2, 2021. Over three days of discussions, the conference was a forum to bring together academic experts, policy leaders and industry professionals to discuss how autonomous systems can be responsibly integrated into the healthcare and security sectors.

With a focus on building trustworthy autonomous systems, the conference covered topics related to both healthcare and security research and identified development areas. The conference addressed a variety of case studies and current research challenges. Delegates presented and discussed the key global issues facing AI development, highlighting the competitive aspects, risks, and opportunities that both nations and organisations will face in the years and decades to come. As a result, keynote sessions, project presentations and workshops have been presented in accordance with the conference scope and discussed challenges, opportunities, and research problems of building trustworthy autonomous systems.

The following report is intended for all those interested in the current challenges and constraints involved in AI development within these sectors. Additionally, the conference features numerous discussions regarding the potential for cross-sector lessons, therefore we welcome readers from the broader AI community who are seeking a concise summary of the global affairs in AI development and implementation.

1. DAY 1: 30TH JUNE

1.1 OPENING REMARKS

The opening session was addressed by **Ms. Emma De Angelis**, Director of Publications/Editor RUSI Journal, RUSI, who declared the Conference officially open and gave an overview of the conference schedule. **Prof Gopal Ramchurn**, Director of TAS Hub, then delivered a welcome talk and introduced a summary of the work in TAS Hub. The opening session then was followed by a creative opening that showed a short film with the title "*Biotech and autonomy collide with the law when a young girl games the system*". The film was written by Luca Vigano, directed by Ali Hossaini, performances by Joan Iyiola and Andreea Paduraru; Photography by Matt Smith, original composition by Keir Vine, coproduced by Ali Hossaini, National Gallery X and Emma De Angelis, RUSI.

1.2 PANEL SESSION: GLOBAL CONTESTS - ARTIFICIAL INTELLIGENCE AND STRATEGY

The chair of this panel session, **Ashlee Godwin**, committee specialist assigned to the UK house of commons and Foreign Affairs Committee, introduced us to the background for this panel session – global contests. Ashlee highlighted the significant questions for global, national security, such as how the global community can manage the threats of AI in the hands of those that wish to do us harm and what it means to be a human working with AI. She outlined what it means to acquire AI capabilities and the power and responsibility that comes with it, quoting Vladimir Putin with "*The global leader of AI will rule the world*". Ashlee emphasised that the UK must lead the AI initiative if human rights are to be embedded into the culture of AI and that national power will no longer correlate as easily to population size. She emphasised that the economy and human rights are leading concerns in the development of AI.

Frank Hoffman, a Distinguished Research Fellow at the National Defence University in Washington DC, was the first panellist to deliver his prepared remarks. Frank began by highlighting how AI will impact many national priorities, including politics, militaries, economics, science, and technology. Whoever leads in AI confers leadership and excels their national security. He emphasised that AI has been noted as one of the significant game-changer to all these aspects.

Frank outlined three major elements of AI development that he sees as being most salient in the years to come:

- Development of socially responsible AI
- Development of AI capabilities
- Competition for AI talent

He stated that the U.S. are committed to developments in line with international laws and values. However, this will inevitably limit AI capabilities, as it cannot be assumed that superpowers like China and Russia will be as constrained by these values. Cyber defence, missile defence and a substantial economic benefit would come to those that invest in AI development. Frank emphasised that the shares of the boom will benefit China primarily due to state control over information. He stated that there are an estimated 11 trillion dollars economic benefits of AI development, and that China will be able to set the standards of operation of AI to control the population.

Further, there would be an impact on unemployment, and we must address socio-political change as time goes on. Frank explains how there is not so much an arms race, but a competition for

human talent. He is confident that Western countries have an advantage because of their universities and companies, as well as their access to entrepreneurialism, and their minimal government interference. STEM graduates are plentiful and provide a substantial economic benefit. Future challenges include securing IP and preserving investment, as well as closing down espionage. It is important not to curtail international cooperation but this will need to be balanced with national development. National strategies are essential. How much does the state mediate AI development? Setting the scene for development? Or providing resources directly?

Dr Kenneth Payne, a researcher in political psychology and international relations at King's College London, was next to deliver his prepared remarks. Kenneth opened his remarks criticising China's AI capabilities. Problems are international, and those that we have are also being faced by China. He disagrees with Frank's previous statements on there not being an arms race of AI. There will be profound effects on fighting power and a lot of uncertainty and fear.

Kenneth outlined the UK's strategy towards AI development by stating that the UK is quite well placed for intense competition in AI. The UK has a rich history and a central university network that can attract talent from across the world. It also has a huge defence budget with opportunities for technology transfer with a good mix of public and private funding in all areas of development. Kenneth outlined some key challenges: attracting talent despite Brexit, and the absence of real tech giants, despite its defence giants China may not have as much innovation power and is still limited by individual capital. In the future, a few things are deemed as being important, building in safeguards, and integrating trust into machines. The UK should lead with creativity and innovation and ultimately with humans at the centre of all concerns. We may need to move away and think about how biotech and AI can be blended.

Dr Pippa Malmgren, economist, author, and technology entrepreneur, was next to deliver her prepared remarks. She opened with the question: can we build something better than the human brain? During the Cuba missile crisis, individuals involved did not want to abandon their families; AI can't match this whole use of emotion and senses. AI is pivotal to future activities, but can we trust it? There are very few data engineers in the defence sector, where sharing data is not commonplace. Pippa goes on to wonder how AI development and roll-out within militaries can or will be managed considering this lack of infrastructure.

The bulk of defence spending is on cyber/quantum computers – yet little is known about this—the most crucial military piece of equipment in the age of AI. AI is about codebreaking at its core: genetic, nuclear, human-personality and behaviour – moving the battlefield from a physical to a digital space. As a seemingly benign example, Tik Tok has shown to be a serious threat to data security. There will be a privatisation of the defence sector over time. AI will likely be outsourced to the likes of Amazon. Companies are becoming very powerful and will play a central role in decision-making. Finally, Pippa ends with her thoughts on the sharp end of the new battlespace: the connection between software and hardware meeting "shareware" – instructions to execute a physical outcome in the real world, and how this will need to be a significant focus for research and development.

Prof Tony Young, practicing frontline NHS surgeon, Director of Medical Innovation at Anglia Ruskin University, was the final panellist to deliver his prepared remarks. He started with a statement: "*The brain consists of 80 billion neurons – quadrillion possibilities*". This network is staggering and difficult to compete with. He states that he might give a different perspective from prior ones because his background in healthcare.

Healthcare is focused on the problem, not the solution. Covid-19 has had a massive impact on the world and has changed how global communities work together on international issues. Covid-19 has affected many healthcare systems. Tony asked: "*How can we use AI to reboot health services across the world?*" In a world that faces an environmental crisis concurrently, equality, diversity

and inclusion must remain central in tackling these issues. The UK has an incredibly competitive health service, holding one of the richest data sets regarding healthcare. How can we learn from it and apply AI to it? The UK can demonstrate which drugs can work and not work to inform international policy and regulation. AI could make use of this dataset. The policy must be built with the population and work with the WHO and other such bodies. The UK must implement, scale and monitor AI in healthcare to ensure that it is both safe and ethical and needs to have the same scrutiny as other advancements in healthcare. AI is meant to assist human action – not lead or decide for iter.

The **Q&A** for this panel session included a question to Professor Young: "*what lessons can the NHS bring to the wider community?"* His reply detailed that the NHS has a broad population of workers and patients/public, creating AI. This can be crowdsourced to create an inclusive culture.

Dr Payne was given the same question. He answered by providing an example of the 'A-level fiasco' and how Covid-19 results outsourced to algorithms. There are major unintended consequences of algorithms that can't always align with human intentions – mistakes will happen. Measures of accountability for these issues must be addressed.

Dr Pippa Malmgren was then asked, '*What are the current legal constraints and regulations for protections in AI?*' Dr Malmgren answered that there are no real laws in place in the new digital world – banks use AI and can see a divorce before the participants too. Drive credit in anticipation. Mental health data? Behavioural data? Trust is an emotion – and that AI is the 'wild-west' of our generation.

Dr Kenneth Payne was asked about his views on EU recommendations for AI development. He answered by stating that the goals and purpose of the government are to promote the health of the population. For the EU, how do we carry on achieving these goals whilst the goalposts keep moving? For example, the right to privacy, legal standards, and retaining the best of this as we go onto the future of shared data.

Dr Frank Hoffman was asked the same question and stated that overall, the EU and US do not like government having information and are incredibly hawkish on AI capabilities and ethics and that the UK probably has the right balance between governmental and individual rights.

Dr Kenneth Payne was asked how the military domain can inform AI development and responded by saying that there is a security dilemma where nations are uncertain about the effect of AI. There is a big gamble to throw in AI to major military systems. AI is developing very quickly, and it's all down to how much governments want to risk. Militaries have been reluctant to change and not mess with things that work. It's fundamentally hierarchical and will require a cultural shift to innovation. The UK will struggle to put these changes into effect.

2. DAY 2 - 1ST JULY

2.1 KEYNOTE ADDRESS: NHS AI LAB- AN UPDATE

The featured speaker, **Eleonora Harwich**, Head of Collaborations at the NHS AI Lab in NHSX, gave a keynote session about the work in the NHS AI Lab. Her work focuses on communicating and engaging with health and care professionals, the public and industry to support and share the work of the NHS AI Lab and the potential of AI. Eleonora started her keynote session by introducing the main goals of the NHS AI lab. She presented three primary goals for the NHS AI Lab:

- Demonstrate the potential of AI-driven technologies for health and care to build understanding among the public and healthcare professionals.
- Build trust and confidence among the public and healthcare professionals in AI, e.g., ethics, transparency.
- Advance regulation steps to ensure that health AI is safe, effective and has equal opportunities across the market. Also, looking for optimising the regulatory process.

Eleonora then described the structure of the NHS AI Lab, which includes three support programmes (strategy and policy, collaboration and engagement and programme management office) and five delivery programmes (Regulation ecosystem, AI ethics initiative, skunkworks, Imaging and AI in health and care awards).

Eleonora introduced the AI Award, run by the Accelerated Access Collaborative (AAC) in partnership with NHSX and the National Institute for Health Research (NIHR). It will make £140 million available over four years to accelerate the testing and evaluation of the most promising AI technologies which meet the strategic aims set out in the NHS Long Term Plan. The Award will support technologies across the development spectrum: from initial feasibility to evaluation within the NHS. The first award has focused on four key areas: screening, diagnosis, decision support and improving system efficiency.

The AI Award is part of the £250 million funding given by the Department for Health and Social Care to NHSX to establish an AI Lab to improve patients' health and lives. Calls for applications for the Award will run at least twice a year through an open competition to identify appropriate AI technologies for support into the NHS. The AI in Health and Care Awards consists of four phases.

- Phase 1: Technical feasibility of the product or concept.
- Phase 2: Prototyping for development and evaluation.
- Phase 3: Real world testing for effectiveness and implementation.
- Phase 4: Adoption into the health system and evaluation of impact.

Eleonora also described the NHS regulation programme to enable a world-leading, transparent, accessible, safe and ethically robust regulation ecosystem for AI-driven technologies in health and care. AI ethics as a responsible approach was also described with a focus on how to counter the inequalities that may arise from how AI technologies are developed and deployed in health and care. Finally, Eleonora discussed the future work of the NHS AI Lab. She viewed that 2022 will witness the launching of the national AI strategy for health and adult social care, in line with the UK upcoming AI strategy. In summer 2021, there will be a launching round 3 of the AI in healthcare awards.

2.2 PANEL SESSION: TRUST IN NON-HUMAN INTELLIGENCE- CAN UNDERSTANDING LEAD TO CONFIDENCE?

The Chair of this session was **Dr Paul O'Neill**, Senior Research Fellow, RUSI. The speakers were **Dr Karen Brady**, **Ardi Janjeva**, **Dr Keith Dear**, and **Dr Michael Nix**.

Dr Karen, Training & Behaviour Consultant at The Guide Dogs for the Blind Association, started her presentation by explaining what building trust in guide dogs for the blind association means. She showed that building trust when the two entities of trust do not speak the same language required extra effort, such as designing communication channels and educational strategy to promote understanding. For instance, she mentioned that they teach people about the limitations of humans, such as overestimating their abilities in doing some tasks. They compare that with highly trained dogs in similar situations. Dr Brady mentioned "*Once trust is built, we start to encourage control of this relationship. We also focus on our later meetings with people at the positive outcomes from the relationship. This helps us to control, monitor and recover from mistakes."*

Dr Keith Dear, Director of Artificial Intelligence Innovation, Defence and National Security, Fujitsu, described that trust between humans is based on ability, integrity and benevolence. However, reflecting these three dimensions on AI will eliminate integrity. Dr Keith mentioned that integrity is not a transferable feature to an AI. Human-AI trust shall be built on understanding the capabilities and limitations of the AI, i.e., what it can perform and what it cannot achieve. Dr Keith presented several recent research findings that show when humans may overestimate AI capabilities and follow their recommendations blindly. When mistakes happen in such scenarios, Human-AI trust is dramatically affected, and it is hard to rebuild trust. Similarly, **Ardi Janjeva**, Research Analyst in Organised Crime and Policing, RUSI, discussed his view on transparency and understanding in applications of intelligence and technology in the criminal justice system in the UK.

Michael Nix, Principal Clinical Scientist -AI implementation, Leeds NHS Trust, defined confidence in AI clinical decision-making as a way of enabling patients and clinicians to place a correct level of trust in every appropriately AI clinical decision, i.e., avoiding cases of overconfidence and underconfidence. Michael presented their work on placing the correct level of confidence in AI clinical decision-making in four dimensions:

- Education: Educate the healthcare worker to understand applicability of AI in their domain, AI expected performance, failure modes and consequences.
- Validation and implementation: We consider workflow integration and clinical assessment of value in situ.
- Communication: We communicate appropriate level of confidence, e.g., explainable AI as an approach. We also communicate risks with patients and clinicians in real world terms.
- Technical robustness. Identify failure cases and predict the estimation of success.

2.3 PANEL SESSION: PROCUREMENT AND ACQUISITION - FUTURE PROOFING FAST-CHANGING TECHNOLOGIES

The first speaker was **Tim Underwood**, who is a professor of Gastrointestinal Surgery and Head of Cancer Sciences Academic Unit at the University of Southampton. He highlighted that UK has the highest number of cases of oesophageal adenocarcinoma cancer in the world. A third of these patients undergo an intensive chemotherapy (and radiotherapy) before their surgery and only 50% will survive five years after their surgery. Despite the complications and side effects of the chemotherapy on patients' health, only 20% respond in a clinically meaningful way to the chemotherapy. A good prediction tool should be available at the time of diagnosis and for every case, inexpensive and scalable across the NHS and other healthcare systems. Diagnosis biopsy

slide meets these requirements besides offering valuable data in the diagnosis of cancer, depth of invasion and it may also contain information about the potential for response to chemotherapy. Professor Underwood and his team used ANNs at the time of diagnosis to predict if the patient will respond to the chemotherapy. The accuracy of this prediction is better than the available methods. Incorporating other input data such as CT scans can add to the prediction accuracy. Professor Underwood thinks that the challenge ahead is to establish trust among his colleagues in healthcare and the patients to be able to rely on these techniques, answer their specific questions and design the treatment according to their preferences. Professor Underwood thinks that his fellow clinicians need to know the details of the input to the model, how it operates and the assurance that the tool is there to support and not to take decision for them in order to trust these automated tools.

The next speaker was **Professor Trevor Taylor** who is a professorial research fellow in Defence Management at RUSI. Professor Taylor highlighted two overlapping areas of activity in defense which are capabilities of development and operation. The simpler the operation context, the simpler it is to enable autonomy at low risk. Most of the decisions in defense are complex which is important when it comes to autonomous systems. Often the immediate progress of the operation (output) is easy to evaluate and accomplish but measuring the consequences of a certain decision (outcome) is a big challenge as it entangles with political point of views and other factors. AI has a wide range of application in defense from intelligence to human resource management, business space, maintenance and repair (e.g., condition-based maintenance), etc. The AI applications in defense can be divided into different levels including analysis, decision prescription and action. Autonomy is rare to be applied to level of action. Given the high level of engagement of private sector in defense consuming about 60% of defense budget, Professor Taylor outlines the challenges of applying AI in defining the boundary of responsibilities for the government, industry and the private sector. Machine learning in defense has to facilitate the selection of relevant data from a large data set, provisioning and curating of data and its ownership, programming, data storage and the protection of the system.

Professor Michael Boniface, the Professorial Fellow and the Director of the IT Innovation Centre at the University of Southampton was the last speaker of this session. Professor Boniface started by mentioning that AI-enhanced digital health is expected to play a crucial role in the management of chronic conditions. However, while the journey for medicines and conventional devices to regulatory approval and then clinical endorsement is well established, the pathway for digital tools and AI is far less. Professor Boniface continues by introducing a project with NHS, called my Smart COPD that focuses on reducing lung function deterioration and increase quality of life by empowering patients and reducing NHS burden. The project is developing a new clinical model for COPD management by prediction exacerbation events using machine learning. Public awareness, perception and involvement, and clinician training throughout all stages of development and operation are essential. There is need for implementation and integration by local trusts and services (as outlined in NHSX Guide to good practice for digital and data-driven health technologies). In the end, the clinical commissioning decisions need to be informed regarding new care models incorporating AI-supported digital health solutions.

During the **Q&A session**, **Philippa Spencer**, the chair of the session asked the panellists regarding the failure of AI in their domains. Professor Taylor pointed out the limitation in the defence sector regarding experimentation. Professor Underwood mentioned that surgeons are the ones who pushed the boundaries of healthcare even though they are quite careful about failure. He mentioned adoption to robotics in healthcare as an example. According to Professor Boniface, there have to be safety measures put in place to ensure patients' safety and the accuracy of prediction of the algorithms. Professor Underwood pointed out that the biases and inherent issues with the data sets regarding the fair representation. For example, the Google AI Skin Diagnosis is based on white skin which will make it very hard to use for other skin colours. Professor Taylor also suggested that the issue of under-representation and bias can also be problematic in other domains.

2.4 KEYNOTE ADDRESS: 'GIVE ME A PING, VASILI. ONE PING ONLY': WHY THE SUCCESS OF MACHINE LEARNING IS DEPENDENT ON EMPOWERED PEOPLE.

Dr Marion Oswald, Vice Chancellor's Senior Fellow at the University of Northumbria, and a practising lawyer specialising in digital usage and data analytics, delivered her keynote speech.

Dr Oswald started with an anecdote, a scene from the film Red October, where the operators track submarines using acoustic software originally designed to track geothermal activity. The algorithm detected a noise as a natural event; however, the human operator was able to deduce that the activity was more akin to a hostile military vessel by piecing together multiple pieces of information. The human supported and explained his findings, convincing the commanding officer to take his story as truth. The moral of the story – don't assume your technology is answering the question you need answering.

Dr Oswald begins to explain how AI could function in a complex world and quotes David Epstein – "*In a truly open-world problem devoid of rigid rules and reams of perfect historical data, AI has been disastrous.*". For example, predictive risk assessments – are they really risk predicting? Since they use group data from the past to predict the future, they are perhaps more accurate to categorise characteristics of the past than what the future holds. AI must evaluate the data used, the data relevant to the question, and the analysis and uncertainties attached to it. There can be errors, and missing information can be amplified if the bottom-up approach isn't used.

Dr Oswald questions whether AI truly works. For example, police forces prefer to be preventive, rather than reactive. A high prediction accuracy at the group level doesn't necessarily mean specificity of the individual. Evaluation of sepsis prediction claimed that the tool identified 183 of 2552 with sepsis (7%). This demonstrated that current AI capacity has very low sensitivity compared to contemporary clinical practice. Overall, many false-positive results can be expected. AI is trying to predict what physicians are already doing and is fundamentally limited to its dataset. For example, the UK's criminal and evidence act states that a person shall be released after a charge unless there are reasonable grounds to consider a risk of re-enactment of criminal activity. While most cases fall within the exclusions, AI may increase suspicions based on the individual's group by creating a risk of general association.

In policing, AI may be useful in tracking organised crime. AI could highlight previous crimes, but it should be scrutinised the same way humans are. Match vs no match algorithms are not so simple. For example, pattern matching threshold values are set by humans to identify faces, so that pattern matching will fundamentally come down to human decision making. Many underpinning rules underneath AI exist and will require humans to inform.

Ultimately, AI development should focus on what the tool is telling us, what is it not telling us, and whether it is relevant to our decisions. Operators and managers need appropriate training and skills. They must make judgments about the relevance of output. Operators need discretion in using AI (if at all) and justify the decision to use it, and management requires a critical approach to AI and the purposes proposed to be used.

2.5 PANEL SESSION: BRIDGING THE GAP BETWEEN LAW AND ETHICS

Professor Chris Watkins, Professor of Computer Science- Royal Holloway, opened the session with **Air Vice-Marshal Tamara Jennings**, a British Solicitor and Royal Air Force officer delivering her prepared remarks.

Jennings states that there is a debate to be had on lethal autonomous systems. If it is all about weaponization, we may miss opportunities outside of this. The military can also free up capacity. AI could benefit not only weapon systems but the repair and maintenance of aircraft. For example,

AI could manage the prediction of maintenance, whether it's deemed as being intermediate or heavy maintenance, and would be incredibly helpful to the military from a logistical perspective. Other logistics solutions include when is best to move equipment and the planning of military operations – hotspots of future events, engagements and population changes and growth because of environmental factors. This will ultimately help protect civilians. AVM Jennings posits the following questions to the audience: are humans any better from a trust perspective? – why? and should an AI system be perfect? Or just better than a human?

Rhodri Morgan, an Electrical Control and Cyber Security Specialist Inspector at the Health and Safety Executive, was next to deliver his remarks. He states that the HSE like to be seen as an enabler, not a preventer.

Health and safety will not change when AI is introduced, the same standards should be met. Incidents could be made by an AI, and when that happens, underlying causes will be analysed to consider whether the causes were foreseeable and how they could have been anticipated. Can it be addressed using humans via addressing workload, equipment, training? Or using AI via the development of good practice, codes of practice and a definition of a global industry standard? During investigation, good practice either has or hasn't been followed. If it hasn't been followed, there is a breach in the legislation.

Finally, **Prof. Subramanian Ramamoorthy**, a professor of trust in autonomous systems at the University of Edinburgh delivered his remarks. He opened by asking: what is different about the introduction of this technology compared to others? There is certainly a shift from the usual "*specification to design*" to "*data to specifications*", in a form of backward design direction. Overall, errors and contingencies are much less criticisable. Extracting exactly the right understanding can be difficult and sub-statements about behaviour remain unclear.

Stakeholder discussions are much harder under these conditions, therefore the relationships in the research and development of AI are challenging. The boundaries between sociotechnical systems and AI is blurred. There are complex interactions, some bottlenecks, and a requirement to negotiate the definitions of terms. The main gaps are insurance and risk, mechanism and market design, negotiation of standards and acceptability. These will be key challenges for the future of AI development.

During **Q&A**, the panel were asked "*Automation is already in play – how can we learn from it?*". The panel answered that we should definitely take the opportunity to exploit current domains that feature AI and continue drafting strategy for developing the technology at its widest context, not just that of war fighting.

The panel were then asked: if an AI system goes wrong, the principle of distinction may be violated, how do we deal with this? The panel responded with an explanation on how a part of the reason why the position is on understanding the black box is to figure out who is responsible and what it can do – Geneva convention may not be addressed if not able to do so. Distinction is a key principle. There is less PTSD and risk with AI, and overall human-harm can be reduced.

The next question asked: "*How do we manage regulations and requirements for competencies?*" The panel answered that this is hard to define and is dependent on context of algorithms. The benchmark is developed to judge the industry and existing regulations will likely be used. The industry benchmark is used to figure out what the competencies are. This does not yet exist, which will pose an issue.

When autonomous systems involve danger, competencies are known, but new applications may not be as easy. For example, in HR, AI algorithms could be used to explore employees' behaviours and anticipate who will leave a company. Who will pick up this responsibility? Technology can be developed for one purpose, and repurposed for another, there is a growing need to regulate its use, and not just at the development stage. From a military perspective the law can be applied to

a variety of contexts. Ecosystem of trust – privacy and governments – fairness and environmental health. Regulators should agree on principles, detailed rules won't help, key principles should be set that are multi-domain. Ultimately, technology evolution is hard to predict.

The panel were then queried with "What about systems that learn?". The panel responded with the observation that we will not be able to directly anticipate the cases and will need to develop a system to anticipate what could happen even when an AI is learning. It's easy to get code onto a device but we must make sure that real-time safety levels are not threatened. For example, Amazon have dismissed employees based on AI, which is cheaper to use than humans looking through data. However, there is a combination of management and AI systems – decisions at scale are an issue and can do a lot of damage. It ultimately comes down to human interaction. The machine is just doing what it's doing. In Amazon's case, the humans have overridden and accepted the risk. Assumably it was a financial decision. Ultimately, being dismissed by a machine threatens the basic premise of human dignity. Ethical principles across the board are essential.

3. DAY 3 – 2ND JULY

3.1 TAS RESEARCH WORKSHOP: HEALTH AND SOCIAL CARE

The workshop on health and social care is chaired by **Age Chapman**, chair of human-centred AI at the University of Southampton, and **Mat Rawsthorne**, a digital health researcher at the Institute of Mental Health at the University of Nottingham. Age mentioned the challenges of integrating autonomous systems into a real-world setting while assuring all parties are happy, including the clinicians, the health staff, and the patients. She introduced sessions that will follow embodied systems, systems with a physical presence and who present a challenge to integrate them to the health care community, and digital systems, where the focus is on building trust and the ethical use of apps and similar systems. Each session is followed by a Q&A session.

TRUSTWORTHY HUMAN-ROBOT TEAMS

As part of the first session on embodied systems, **Nicholas Watson**, Associate Professor of chemical engineering from the University of Nottingham, presented the TAS-Hub Agile project called Trustworthy human-robot teams. The project is highly multi-disciplinary, with computer science, engineering, social sciences, and law researchers, and collaborates with Agri Forwards and Intuitive Surgical as industrial partners. The project concerns two case studies of human-robot teams, a human-robot cleaning team and a human-robot surgery team. The project focuses on user studies to better understand how trust can be generated within the human-robot team and from the broader community towards the human-robot team. The project also aims to develop and test new methodologies for improving and verifying the effectiveness of human-robot teams.

Marise Galvez Trigo, PhD candidate in computer science at the University of Nottingham, presented the first case study on human-robot cleaning teams. Marise stresses the potential of disinfection robots as they can be used in the food industry, health care, and schools, especially in the current COVID-19 pandemic. In particular, she mentions that UVC-robots emit UV rays to deactivate viruses, fungi, and bacteria, and therefore could be used in many of these settings, although more research is needed. The project will investigate trust within teams and towards teams with questionnaires, user studies, and interviews. The effectiveness of the UVC-robot will be assessed by placing UV sensors around the room and testing the doses of UV received**.**

Dr Sylvaine Tuncer, Research Associate in Sociology at King's College London, presented the second case study on human-robot surgery teams. In robot-assisted surgery, the robot mediates the surgeon's actions. Communication in robot-assisted surgery is mediated through loudspeakers and video footage of the patient's body. This raises the challenge of situational awareness, namely, can the surgery team communicate efficiently and agree on what is going on? The case study will investigate trust towards the team by interviewing patients and a variety of staff (management, administrative, nurses, anaesthetists, etc.). The case study will also investigate trust within the operation theatre by video recordings, observations, and breaching experiments in which a simulation is performed to see what happens if the robot fails at a particular part of the task.

TOWARDS RESILIENCE OF AUTONOMOUS PHYSICALLY ASSISTIVE ROBOTS

Prof. Sanya Dogrammadzi, Professor of medical robotics at the University of Sheffield, presented work that is part of the Resilience Node and is coordinated by the University of York.

Sanya mentioned assistive robots which can serve a variety of purposes in homes and health care, including helping to maintain quality of life for people with physical impairment, rehabilitation and recovery, and reducing labour shortages. In conjunction with the Internet of Things (IoT), assistive robotics can also be used as diagnostic tools for caregivers. This field presents some

challenges that are actively being investigated. Adaptability is one of the important challenges: as no two users are the same and the user may change preferences or needs over time, an assistive robot must be able to adapt online to user requirements. Safety is an important concern and various digital technologies are being explored, including teleoperation, cybersecurity, IoT and safety sensors.

There are a variety of scenarios, and not all of them can be anticipated. Interaction complexity, autonomy, and safety are intimately linked, and hazard analysis clarifies the link. The environment was analysed into different agents, objects, terrain, ambient conditions, etc. Levels of autonomy can vary from manual supervision to completely unconstrained in open environments. Sanya shows a robot assisting a human in dressing, with one HHI study investigating disambiguation of deictic language and one HRI study focusing on distraction and safety during dressing.

KASPAR EXPLAINS

Prof. Farshid Amirabdollahian is a Professor of HRI at the University of Hertfordshire who presents the Kaspar Explains project. Kaspar is a humanoid robot that looks a bit like a child and provides autism education, particularly with the concept of causality. The robot offers causal explanations for educating on desired behaviour in social interaction (e.g., learning not to hit Kaspar by letting Kaspar discourage certain tactile behaviours). The project works with various partners, including Compusult, GMS, and a local school. The technology is expected to have commercial potential and will be presented at multiple policy and public events.

WOULD YOU TRUST A ROBODOC?

Dr Jonathan Ives, Professor of Empirical Bioethics at the University of Bristol working within the TAS Functionality Node in the University of Bristol, considers trust, trustworthiness, and how they develop when an autonomous system has an evolving functionality. Jonathan stated that trust is a belief that the object or agent will behave as expected and is mainly learned via induction: one observes a few successes and generalises; however, the generalisation may fail. We trust toasters or cars because they work, even though we don't know how they work, and we trust doctors because they are highly trained. Evolving functionality may undermine trust as it is unpredictable. Understanding it better might help restore trust, but autonomous systems tend to be complex and opaque, making that difficult. For the case of health care professionals, we trust them because there is a legal framework around the profession, because there is a lot of evidence that can be consulted in line with the practice of the profession, and because the health care professionals keep demonstrating competence and ethics. Are autonomous HCPs that different? They do not have a professional association, have not been verified in the same way, and lack liability.

AN OPEN LABORATORIES PROGRAMME FOR TAS (OPEN-TAS)

Prof. Tony Prescott, Professor of cognitive robotics at the University of Sheffield, presents OPEN-TAS, a pump-priming project to make TAS open to the general public.

The project aims to set up user-friendly, remote robot telepresence, using a VR headset, in the Open Lab, where science in the field is demonstrated. One of the key motivations to do this is that to trust the general public, there needs to be transparency, and open science is a great way to do this. The project leads to two spin-outs. First, the company CyberSelves focuses on understanding how immersive technologies impact society. Animus is the key technology used in CyberSelves. Animus is a body-agnostic approach to robotics, allowing code for different robot platforms and providing low-latency communication. Animus has been used for telepresence in a robot body via VR, as demonstrated in the Teleport app developed within the CyberSelves

project. The Teleport app allows you to select one robot of a few options to which you can teleport (e.g., the Tiago robot, MiRo, Pepper), and then you can view the world from its perspective and control it remotely. This will be useful for variable autonomy, i.e., when the robot is stuck, you can help it. That is, one of the uses for telepresence is when the robot cannot be fully trusted in its autonomy.

TRUSTWORTHY AUTONOMOUS SYSTEMS TO SUPPORT HEALTHCARE EXPERIENCES

Liz Dowthwaite, Research Fellow in Social Psychology at the University of Nottingham, presents the project Trustworthy autonomous systems to support healthcare experiences. The project focuses on user experiences with health monitoring systems. These days we have smartphones, smartwatches, smart toothbrushes, and even smart mirrors, with various apps aiming to improve habits related to personal care, exercise, sleep, brushing teeth, etc. This project focuses on smart mirrors to monitor and support personal care and wellbeing. Its functionalities range widely: monitoring posture and facial expressions for possible psychological guidance; providing reminders of daily care routines such as tooth brushing or washing face; and alert family or medical professionals in case of emergency. This can be especially useful for people with dementia, people who had a stroke, people with multiple sclerosis, or otherwise vulnerable people. The technology is powerful, but more research is needed. These data are pretty sensitive, and the user experience requires some fine-tuning through user feedback. Also, the use of such apps should be considered within the more comprehensive network, not just the person using it but also the family members and health care staff.

EMBEDDING SOCIAL SCIENCE INTO UPSTREAM ENGINEERING RESEARCH: INCLUSIVITY AS A PATHWAY TO BUILDING TRUST

Dr Stevienna de Saille, Research Fellow in sociology at the University of Leeds, presents the project on embedding social science into engineering research, particularly on the acceptance of robots to perform care functions. Surveys and questionnaires do not necessarily capture everything that people think. So why not try to capture what people imagine? This projects explores how people tell stories about robot carers using a Lego toolkit. For example, some stories from the participants were nurses getting in the way of robots. Such studies allow us to investigate and discover some underlying themes on robots in care functions, in particular, trust, robots as a companion, convenience of automation, dependence on the technology, and the control problem of whose needs are served by the robot (the human user or the robot itself, for example).

UNDERSTANDING TRUST AND PUBLIC ACCEPTANCE OF DIGITAL CONTACT TRACING IN THE UK

Dr Joel Fischer, Associate Professor of HCI at the University of Nottingham, talked about contact tracing. Contact tracing apps have agency: decisions to instruct the user to self-isolate are based on algorithms. More specifically, such apps detect whether you have been in close proximity to others and if so, advises you too self-isolated. The project studied differences between groups (e.g., those with greater risk factors) and the study found that:

- 50% downloaded
- 27.4% did not participate
- BAME and older adults have lower participation
- 40% do not want to be tracked
- 30% do not think it is effective
- 30% do not trust the app

Users that deleted the app or did not download the app did not really trust the app. They had a variety of concerns: how the data is used, did not understand how the app worked, thought regulations were insufficient, or that the app was unreliable or did not do what it claimed. Trust issues seem to be exacerbated among vulnerable people, leading to the paradoxical situation that those who need help seek it the least.

3.2 WORKSHOP 2: DEFENCE & SECURITY – PART ONE

Stuart Middleton, Lecturer in Computer Science at University of Southampton**,** and **Alec Banks**, Senior Principal Scientist at Dstl affiliated with the University of York, presented the Defense and Security workshop.

SECURITY ASPECTS IN DEFENCE

Prof. Gokhan Inalhan, BAE Systems chair, presented the security aspects in defence. He discusses, in particular, improving autonomy, providing more mobility across the air, and the connectivity of such systems with themselves and with human operators. The research takes place within the Security Node, focusing on usage, operations and users of autonomous systems.

One particular concern is to provide runtime adaptivity to (perceived) attacks, such as taking out GPS or pixel attacks. The solution proposed by Prof. Inalhan is to combine interpretability, continual verification, and adaptability. Applications include adaptive flight control systems where real-time adaptation is required.

TRUSTWORTHY SWARMS IN EXTREME ENVIRONMENTS

Dr Mohammad Divband Soorati, Research Fellow in swarm robotics at the University of Southampton, presented Trustworthy Swarms in Extreme Environments, an Agile project in the TAS-Hub.

Swarm robotic systems have many benefits. First, they have robustness, decentralisation, redundancy, simplicity and distributed sensing imply systems can detect errors within themselves and keep performance when few robots fail. Second, because they are composed of many small elements, one can reconfigure them easily to adapt to task requirements. Third, changing the number of robots in the swarm typically does not drastically affect performance.

However, urgent decision-making and having a global view of the environment are difficult with such systems. Therefore, Dr Soorati proposed that humans can help swarms make urgent decisions and support their understanding of the environment. Such human-swarm systems are difficult to set up because of the complexity of the swarm, and it is not always clear when the swarm can be trusted in its autonomy.

A use case of the algorithms and tools developed in the project are searches and rescue in hurricanes. In this use case, one needs to identify locations of survivors, evaluate their health status, and evacuate them with UAVs. During this time, the human guides the exploration process.

In applications such as these, user interfaces and explainability are vital concerns, and data-compression techniques will also be used to ensure that the constraints in communication are respected.

SWARM ENGINEERING ACROSS SCALES

Dr Sabine Hauert, Associate Professor of Swarm Engineering at the University of Bristol, presented Swarm Engineering across scales. Swarms have distinguishing properties such as

robustness of swarms to failures, namely if one agent fails, the swarm does not fail, and to swarm size changes, namely if swarm size increases, the swarm maintains performance.

Dr Hauert discusses a variety of applications. There is great enthusiasm to design swarms for applications such as fire-fighting, but there is still uncertainty and lack of trust in their performance. Logistics in smaller, messy warehouses (as opposed to the vast, highly organised warehouses in the likes of Amazon) may also be one domain where swarms can help, since most automated systems work well in structured environments but not in chaotic environments, while swarms can do well in the latter. Similarly, cloakrooms might be operated by swarms to help people get their jackets back. She also discusses applications with nano-particles. For example, machine learning can be used to engineer useful swarms of nano-particles, changing their sizes, shapes, and charges with different molecules for cancer treatment. Last, she discusses the design of devices for wound-healing.

UNCERTAINTY-AWARE MACHINE LEARNING FOR TAS

Liyudmila Mihaylova, Professor of Signal Processing & Control at the University of Sheffield, presents uncertainty-aware machine learning for TAS.

She presented first the project SIGNetS, Signal and Information Gathering for Networked Surveillance. With large sensory networks providing heterogeneous sensory information, how can their data be integrated? The research will focus on resilience and trustworthiness (theoretical, software, algorithmic), fusing data from heterogeneous sensors, and linking technological aspects with ethics and other human factors.

She then highlighted that robustness to changes is essential, as, for example, Tesla automated driving system had some recent failures. Systems must have uncertainty awareness, be aware of limitations in data, models, and environments, and perform continual learning and be explainable. Prof. Mihaylova then presented the various methodologies she is exploring, in particular, recursive Bayesian methods, nonlinear regression, unsupervised activity understanding (for example to understand traffic), robustness to adversarial attacks, and formal verification with confidence intervals.

3.3 PANEL SESSION: BIOLOGICAL AUTONOMY: CAN MACHINES COME ALIVE?

The Chair of this session was **Professor Luca Viganò**, Vice-Dean and Head of Cybersecurity Group, King's College London. The speakers were **Dr Ali Hossaini**, Senior Research Fellow at King's College London and Director of National Gallery X, **Prof. Denis Noble,** Emeritus Professor of Biology at Oxford University, **Prof. Ana Soto,** Professor of Immunology at Tufts University, **Prof. Ray Noble,** Emeritus Professor of Biology at University College London and **Professor Carlos Sonnenschein**, Tufts University School of Medicine.

The panel raises various possible definitions of what it means to be "*alive*", including agency, motility, self-awareness, self-preservation, self-correcting, autopoiesis, creating its own rules, choosing among options, creating novelty. **Dr Ali Hossaini** provided a thermodynamic/energetic definition, namely an agent is alive if it is auto-katalytic with various work-cycles (i.e., it seeks out its own fuel). Contrast this to viruses that have an inert reproductive code and are katalytic. **Prof. Ana Soto** Professor of Immunology at Tufts University, states that mathematicians define life as the creation of negative entropy. **Professor Carlos Sonnenschein** raised the question of what life is? And how to differentiate between life and no-life?

In another topic, the role of biological autonomy in the creation of AI was discussed. **Prof. Denis Noble** stated that the stochastic combination of molecules to gain order we see in biological

organisms might be essential to AI. **Prof. Ray Noble** noted that there might be risks introducing behaviourism in biology, where no agency is assumed, stating, "*we are not machines*".

Finally, they discussed catastrophic risks associated with biological autonomy. Even low-level intelligence "*without a plug*" could potentially be extremely dangerous. Intelligence and even super-intelligence are a tool, but autonomous life of different sort would be a threat. Similarly, boundary cases such as new viruses are also a great threat, despite their complete lack of intelligence.

3.4 TAS RESEARCH WORKSHOP: DEFENCE AND SECURITY: PART TWO

ASSURING MACHINE-LEARNING IN AUTONOMOUS SYSTEMS (AMLAS)

Collin Patterson, Research Fellow in Computer Science at University of York, presented AMLAS (a project which analyses safety case patterns to define safety analysis for a particular component in an ML system). The analysis methodology starts with a safety requirement, provides learning and verification assurances, and finally checks how the system would work in deployment. For example, for detecting a pedestrian crossing the road, the ML requirements specify position prediction accuracy (e.g., up to several pixels) and establish robustness to lighting conditions, pose, etc. One then has to analyse the relevance to the system used in practice and the completeness and how well the data covers the domain.

PARTICIPATORY ASSURANCE METHOD

Dr Christopher Burr, Research fellow at Alan Turing Institute, started by introducing the participatory assurance method. He mentioned that providing assurances to issues such as explainability and data privacy issues requires a participatory approach that includes affected stakeholders within the design, development, or deployment of the respective technology. This project collaborates with a range of stakeholders to understand what issues matter most to them and how to develop a justifiable method of assurance that helps promote trust and confidence. Then, Dr Christopher proposed the main challenges for AI tools following design, development and deployment framework. For instance, Dr Christopher identified transparent procurement and sustainable infrastructure as the main challenges for the development phase. Then, Dr Christopher explained the argument-based assurance concept as a process of using structured argumentation to provide assurance to another party (or parties) that a particular claim (or set of related claims) about a property of a system is warranted given the available evidence. Finally, Dr Burr introduced a participatory assurance method that helps to ensure that any significant issues are identified, and any harms or risks are mitigated as early as possible in system deployment.

REGULATORY CHALLENGES AND TOOLS TO ADDRESS THE FALLIBILITY OF AI-BASED SYSTEMS

Prof. Subramanian Ramamoorthy, Professor of Robot Learning and Autonomy at the University of Edinburgh, studied how perception errors in autonomous systems can have profound effects. A typical example is pedestrian detection. AI is not robust to locate errors.

He then presented some of his research, starting with Vignette, a probabilistic environment specification language which allows to generate lots of different scenarios according to a particular specification, with improved explainability. This was studied in a pick-and-place robot case study. Second, he presented constrained learned models with human guidance. Based on

high-level specifications (e.g., domain, regulations) on the environment in addition to the learning on the data, we can then get constrained models. A case study is UAV autonomy.

CONSENT VERIFICATION IN AUTONOMOUS SYSTEMS

Dr Inah Omoronyia, Lecturer in Software Engineering at the University of Glasgow, presented consent verification in autonomous systems. He discussed developing automated reasoning techniques for verification. Namely, we wanted to replace deterministic systems with automated reasoning, resulting in privacy-respecting autonomous systems that consider data regulations. These would replace a data protection officer by 1) accurately identifying personal data, 2) ensuring there is no emergent sharing data, and 3) the purpose of data processing remains unchanged. Autonomous systems cannot guarantee these, so the research must address three questions: 1. to automatically determine whether data is personal; 2. to automatically determine when to seek consent, and 3. to automatically detect whether it is in the best interest.

SAFESPACESNLP

Jeremy Clos, Research Fellow in Computer Science at University of Nottingham**,** presented SafeSpaces NLP, a project to make online forums safer for children. To solve this, they used behaviour classification using NLP. Dr Clos noted that moderating posts in Kooth (and other platforms) takes too much time, so an automated approach to moderation makes sense. The traditional approach is to train on a vast dataset but not generalise to slightly different tasks. Putting humans in the loop can help.

Dr Clos then presented various approaches in socio-technical NLP, including interactive sense-making based on visualisation, explainable AI using an explanation model to analyse failures and then retrain, active learning, adversarial training, and meta-learning/few-shot learning.

The project combines insights from multiple disciplines, including computer science, linguistics, criminology. As stakeholders give feedback, Dr Clos believes there will be more and more trust over time. Activities in the different work packages include

1. Annotation of datasets / interviews with moderators.
2. Deep learning algorithms.
3. Evaluation in Kooth case studies plus standard benchmark datasets.

3.5 CLOSING REMARKS

Prof Gopal Ramchurn, Director of TAS Hub, opened up the closing session and thanked everyone, including speakers and attendance. The closing remarks session included summarising key ideas that have been discussed during the conference with the workshop chairs **Dr Stuart Middleton, Dr Alec Banks** and **Dr Mat Rawsthorne**. Gopal also invited **Dr Ali Hossaini** and **Ms Emma De Angelis** to present their views about the conference.