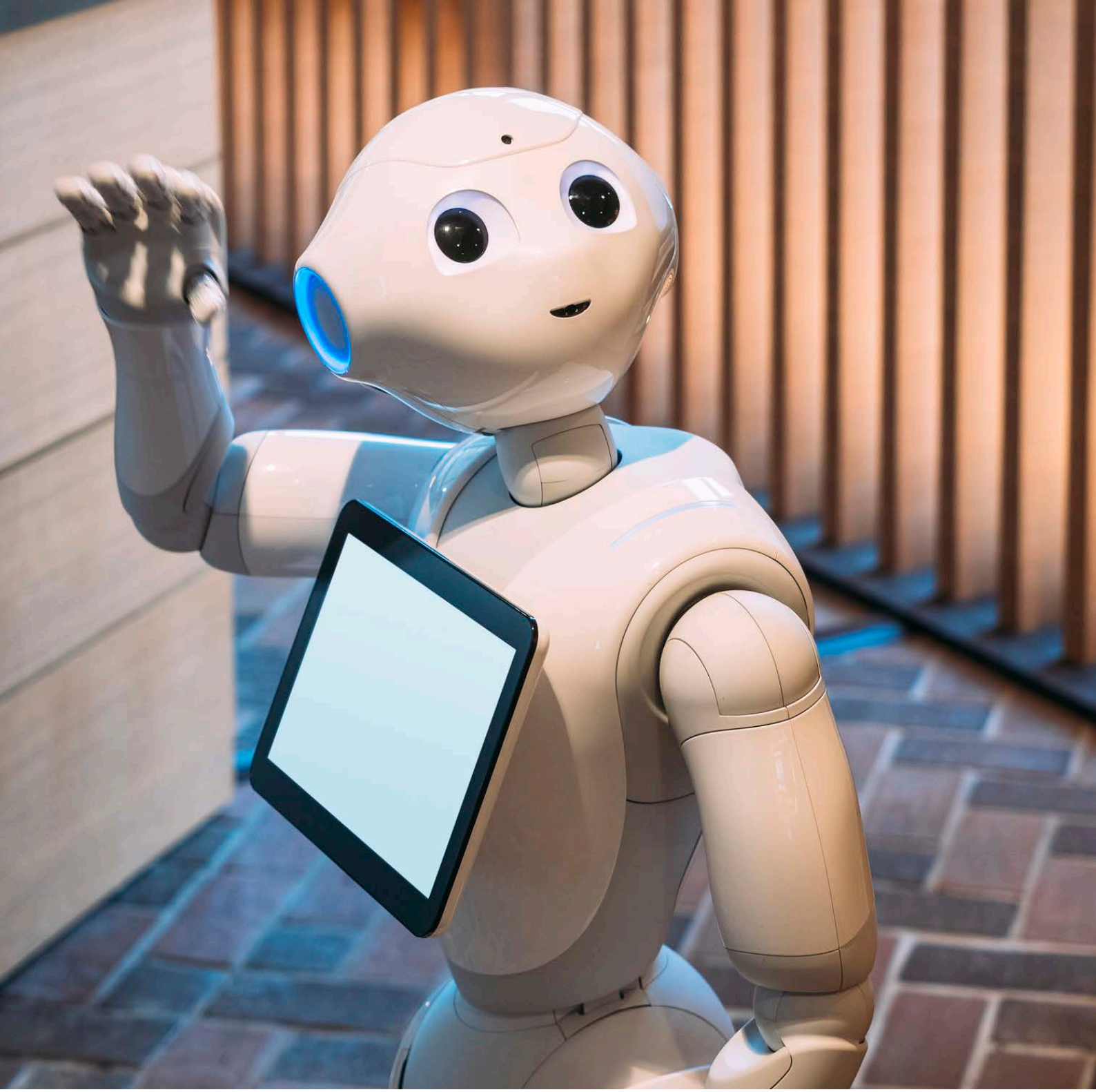


UKRI Trustworthy Autonomous Systems Hub

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?

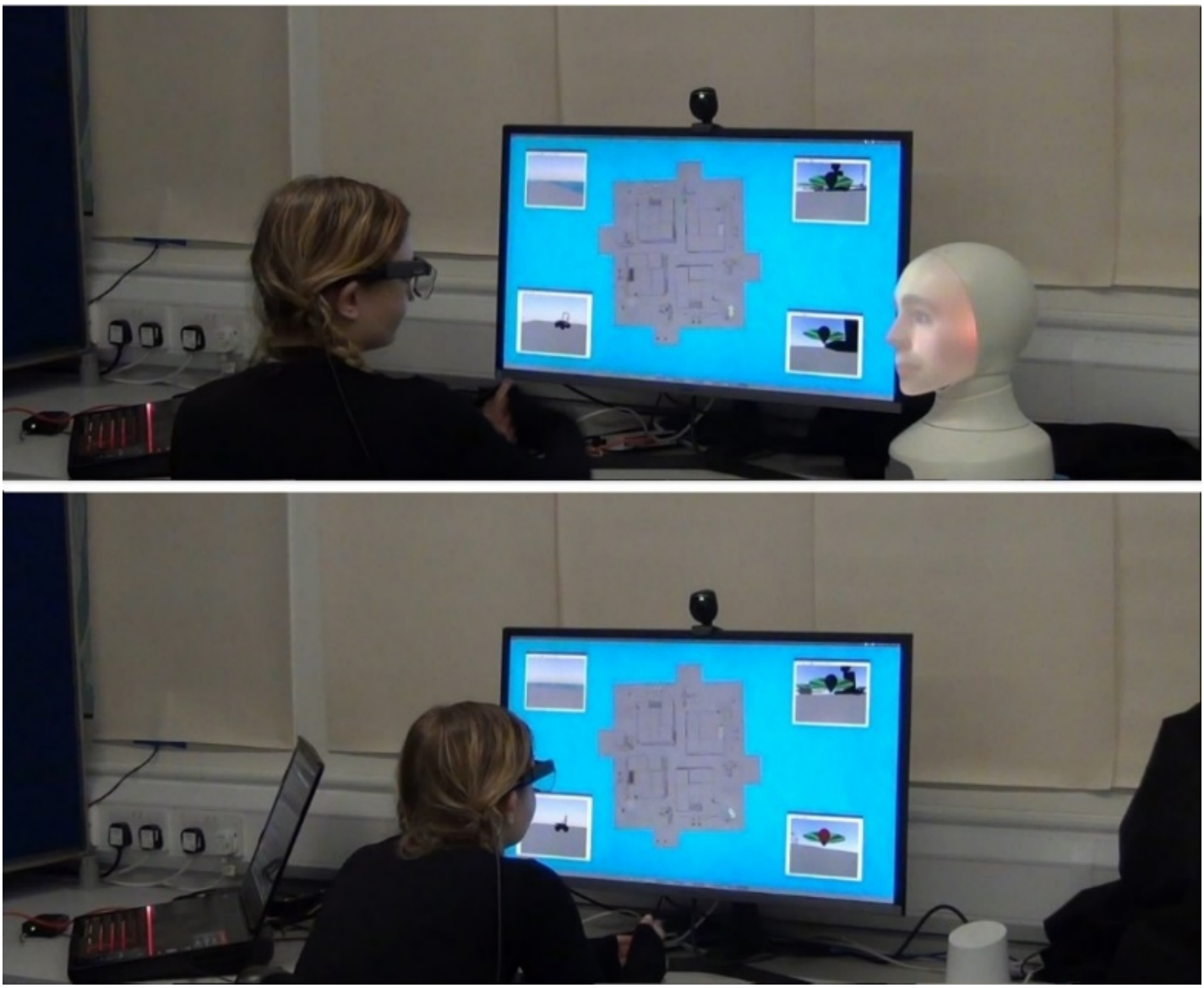
Exploring human-robot trust

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).

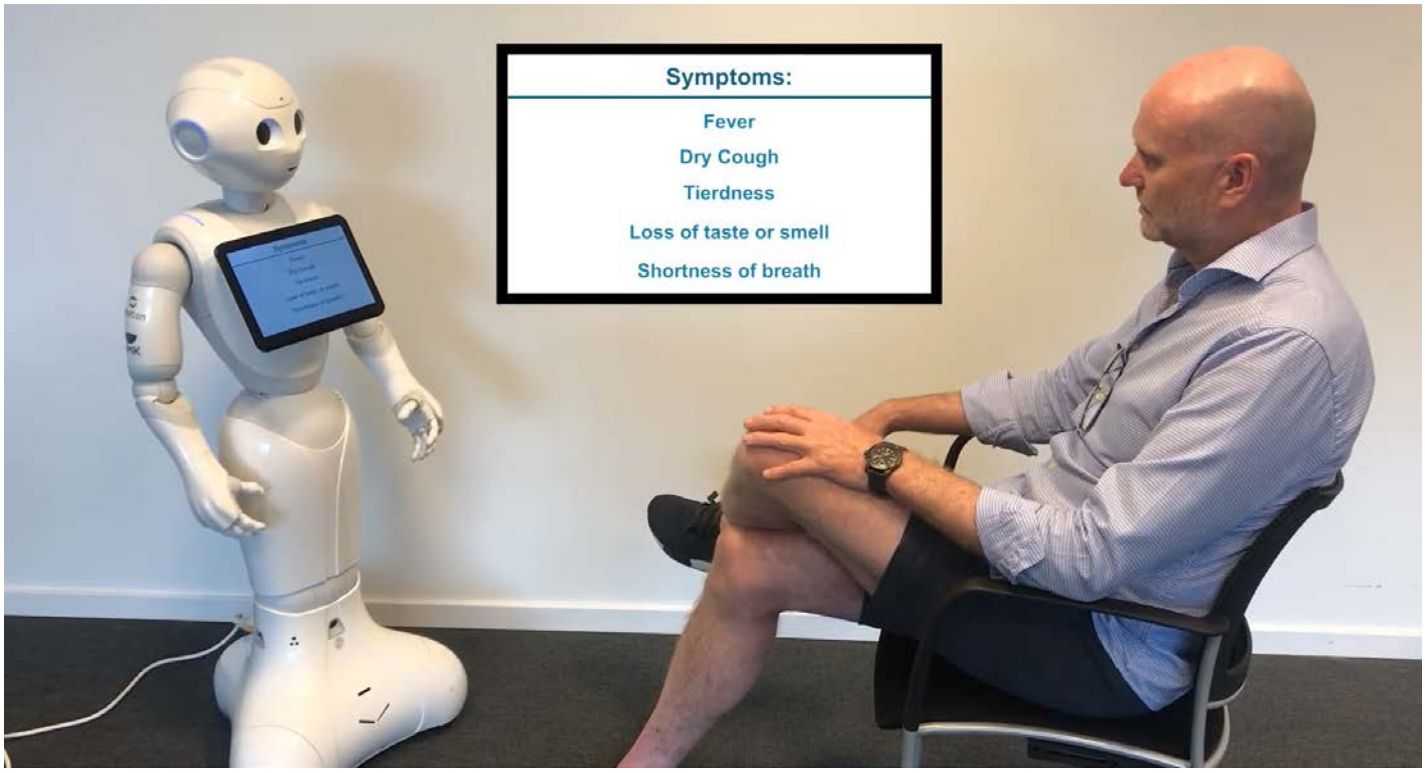
Another investigation used an online maze to explore how trust forms between humans and robots while following recommendations. A user with the goal of escaping a maze could either accept or refuse the robot's help. The study explored how our trust and behaviour was affected when interacting with three 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.

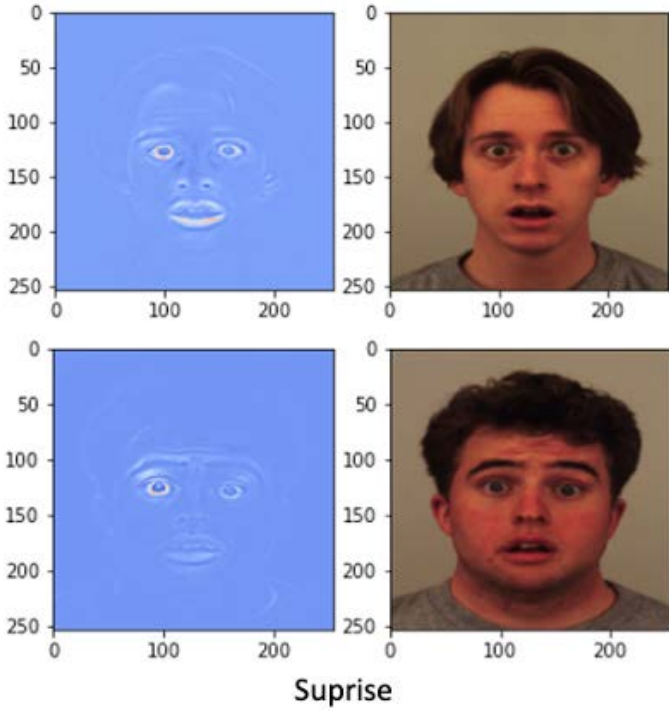
However, there is a balance to be struck and too much communication has the opposite effect. If a robot constantly requires too much feedback or asks too many questions, we tend to lose trust in it.

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