Australian National University

Australian National University
Institute for Space

Australian Government
Defence

Policy Brief

Trustworthy Autonomy and AI in Safety-Critical Domains

Overview

In June 2024, the Australian National University's School of Engineering and Institute for Space (InSpace) collaborated with Defence Science and Technology Group (DSTG) to deliver the first Theory to Practice workshop, focussing on safe and trustworthy autonomy and AI. This policy brief summarises the outcomes of the workshop and presents four recommendations moving forward.

Workshop aim

The aim of the workshop was to explore how theory can be translated into practice for the safe and trusted implementation of autonomous and AI-enabled systems in safety critical applications.

Workshop participants

The workshop participants consisted of members from Defence, Defence industry and academia. The audience was focussed on Defence contexts and applications of autonomy and AI in safety-critical domains and applications.

Disclaimer

The content within this brief is a summary of the opinions expressed at the workshop. Where additional content has been added, relevant references have been provided.

Acknowledgements

This policy brief was written by Dr Zena Assaad with the support of the panel members from the Theory to Practice workshop: Adam Hepworth, Benedict Lyons, Chris Kourloufas, Elizabeth Williams, Hanna Kurniawati, Luke Grassy, Mel McDowall, Rachel Horne and Thomas Graham.

Safety of Autonomy & Artificial Intelligence

Autonomous and artificial intelligence (AI) capabilities are becoming increasingly more prevalent in technology applications, specifically those in safety-critical domains. These capabilities bring with them a number of unique safety challenges which require a shift in how safety is understood.

Autonomy refers to the capacity to operate independently of direct control.[1]

Artificial intelligence refers to systems which analyse large amounts of data to find patterns in that data to inform outputs, which can be adjusted via a learning process.[2]

A definition of safety which fittingly describes this evolving landscape is "freedom from harm or unintended outcomes".[3] The increasing complexity of autonomous and AI-enabled systems coupled with the ubiquity of their scale requires a much broader consideration of possible harms. This has resulted in a more socio-technical understanding of safety, which includes concepts such as trust. The concept of trust is similar to that of safety. Safety is not a capability of a system, rather it is a standard of operation. Measures, procedures and boundaries are implemented to assure and demonstrate the safety of a system when in operation.

Harm refers to human experiences which result in some kind of negative impact (injury, loss, damage, distress, etc). These impacts can include, but are not limited to, physical, psychological, social, cultural, autonomy, economic and political.[4]

Unintended outcomes refer to the link between an action or inaction and an outcome.[4]

Similarly, for autonomous and AI-enabled systems, trust can also be thought of as a standard of operation. The critical question which emerges is, what measures, procedures and boundaries can we put in place to demonstrate that a system is trustworthy?

ANCAP safety ratings are not mandatory. ANCAP is an independent non-regulatory consumer information organisation. The ANCAP safety ratings have elevated the safety of cars on the market across Australia and New Zealand because the ratings create competition among manufacturers, encouraging them to improve safety for a commercial advantage. These ratings are also useful to consumers because they are a clear and easy to understand benchmark for safety.[5]

Safety assurance is a legally mandated requirement of safety-critical systems and can play a role in shaping trust dynamics. However, mechanisms for assuring trustworthiness of such systems are more extensive, and are not yet legally mandated or incentivised. The incentive to assure trustworthiness of these systems diminishes in absence of legal mandates, relying on the development and acceptance of norms to deliver assured trustworthiness. As the promulgation of regulations has historically lagged behind the advancements of technology capabilities, non-legal approaches to encouraging the assurance of trustworthiness will need to be explored.

Complexity and context are two critical factors impacting both the safety and trustworthiness of autonomous and AI-enabled systems. Complexity is an emergent property, arising from both technical complexity and the complexity of humans operating alongside these systems. Context is what shapes the finer details of an operating environment. This directly impacts how a system operates, responds to and changes within that environment and how its actions are likely to be perceived.

While there is increasing evidence demonstrating the significance of trust for autonomous and AI-enabled systems, particularly in safety critical domains, the lack of consensus on what this term means and how it can be captured are stagnating the translation of theory to practice.

Understanding Trust

Trust is a multifaceted concept which is understood and defined differently across different industries and areas of academic research. This dissonance has created a number of challenges for capturing trust, particularly in safety-critical systems.

Differing conceptions of trust have led to a diversity of language used to describe this term. This has created greater levels of ambiguity and subjectivity. It has also led to a lack of specificity in how trust is captured in practice.

Trust is closely linked to public social licence. A general lack of trust from the public in autonomous and AI-enabled capabilities creates barriers for the acceptance and adoption of these capabilities. A lack of trust also influences how people choose to interact with a system, which can lead to greater safety implications. While the promulgation of regulations is lagging well behind the development of autonomous and AI-enabled systems, public pressure may actually aid in evolving regulation faster.

A working **definition of trust** adopted within this policy brief: "*confidence in the reliability of a system when used in the intended operation of use.*"[6]

While there is yet to be a consensus on the definition of trust, there is a general acceptance that trust is a human concept. Machines do not hold the capacity to trust. Therefore, measures for encouraging or enabling confidence in the reliability of a system must be considered at the human level.

Under trust - Having little to no confidence in a system. In these situations a human may ignore or misuse that system, foregoing its capabilities and purpose.[7]

Over trust - Having too much confidence in a system. In these situations a human may not detect errors, malfunctions or incorrect outputs from the system because they assume it will always produce the correct or appropriate output.[8]

Because regulatory initiatives are proportionate to risk, a risk-based approach to regulating trust in autonomous and AI-enabled systems may be a strategic pathway for managing trust. Particularly when considering how interlaced trust is with safety.

Anticipatory regulation also presents an opportunity for proactively and iteratively developing regulations; however, there are few good examples of anticipatory regulation in practice. The aviation industry has attempted to embrace anticipatory regulation; however, the pace of technology advancements and the scale of their applications has made this endeavour more challenging and complex.

Additionally, the rapid pace of development and far-reaching scale of implementation creates an unstable environment for trust perceptions. One accident or one unintended outcome can eliminate trust almost immediately. The social licence required for these systems is closely tied to human perception, making the problem of establishing and maintaining trust a socio-technical one. Technical approaches to trust may address system level considerations; however, they will not capture the foundational human element of trust.

Theory to Practice

Understanding trust and its implications for the safety of autonomous and AI-enabled systems in safety-critical domains is an ongoing area of research. The challenges of varying contextual applications, the spectrum of technical capabilities and the requirements that come with safety-critical systems have stagnated the transition from theory to practice.

Different domains have progressed at different rates, with some facing more barriers than others. For safety-critical applications of these technologies, the progress has been far slower because of the greater levels of risk associated with these applications.

Engaging stakeholder groups across academia, Defence and industry is both a strategic opportunity and a key challenge. Strategically, leveraging the expertise across these stakeholder groups would enable more robust outputs; however, differing priorities across these groups removes cohesion.

Establishing a common mission across these stakeholder groups can aid in bringing the different conversations together. In the case of autonomous and AI-enabled systems, the common mission is capturing trust in practice in safety-critical domains.

To achieve this, there would first need to be agreed upon definitions to ensure a common contextual benchmark. There would also need to be elements of participatory or co-designed initiatives to ensure the needs of each stakeholder are reflected in outputs.

While different user groups will have disparate needs, the current ambiguous and contested safety and trust landscapes of autonomous and AI-enabled capabilities is creating large amounts of discourse debating semantics, which is stagnating progress on safe and trusted implementation in practice.

Broadly speaking, safety is a concept which holds a common understanding across different disciplines. The specific approaches or measures for how safety is assured across these disciplines is what changes. A similar approach can be taken with trust. Developing a broadly accepted understanding of trust and complimenting that with domain specific measures or practices for achieving and assuring trust.

Legacy practices are also hindering progression in how safety is managed and assured for these systems. Safety is traditionally a very predictable and deterministic process, with little room for uncertainty. The nature of autonomous and AI-enabled technologies challenge this status quo.

In order to translate theory to practice, some benchmarks need to evolve or be shifted to meet the demands that come with autonomous and AI-enabled technologies in safety-critical domains.

Recommendations

Establish a common vocabulary for trust in specific contexts

Measures around trust cannot be developed without clarity on what trust means. In the case of safety-critical systems, trust ***must*** be clearly defined. The subjectivity of this term has stagnated consensus on how it should be defined, with the nuances of context influencing discourse. It is therefore recommended the vocabulary around trust be defined subject to specific contexts and applications.

Develop a taxonomy for autonomous and AI-enabled capabilities in specific contexts

Autonomous and AI-enabled systems vary across a broad spectrum of technical capabilities. As such, one blanket approach to all capabilities which are labelled autonomous or AI will ultimately result in disproportionate measures for many of these capabilities. It is therefore recommended a taxonomy for autonomous and AI-enabled capabilities in specific contexts be developed. The taxonomy should outline the capabilities and limitations of autonomous and AI-enabled systems in specific contexts.

Improve general education on autonomous and AI-enabled systems to encourage calibrated levels of trust

The combination of the broad spectrum of technology capabilities and evolving public narratives around these capabilities has fuelled many inaccurate perceptions and understandings on what these systems are and are not capable of. False perceptions and inaccurate understandings can lead to miscalibrated levels of trust (i.e. over trust or under trust). It is therefore recommended general education around autonomous and AI-enabled systems be improved, with particular focus on the capabilities ***and*** limitations of these systems.

Promote participatory and co-designed outputs within a multidisciplinary ecosystem

The broad and diverse implications of autonomous and AI-enabled systems influence a plethora of considerations including psychological, social, cultural, economic, legal, political, and many more. Addressing this requires a multidisciplinary approach. It is therefore recommended that participatory and co-designed outputs be promoted to encourage a multidisciplinary ecosystem of work.

References

1. Lyons, J. B, K Sycara, M Lewis, and A Capiola. “Human–Autonomy Teaming: Definitions, Debates, and Directions.” *Frontiers in Psychology* 12 (2021): 19–32.

2. Glikson, Ella, and Anita. W Woolley. “Human Trust in Artificial Intelligence: Review of Empirical Research.” *Academy of Management Annals* 14, no. 2 (2020): 627–60.

3. Leveson, Nancy G. *An Introduction to System Safety Engineering*. Cambridge, MA: The MIT Press, 2023.

4. Aven, T. “Risk Assessment and Risk Management: Review of Recent Advances on Their Foundation.” *European Journal of Operational Research* 253, no. 1 (2016): 1–13.

5. Newstead, S. & Scully, J. "Potential for Improving the Relationship between ANCAP Ratings and Real World Data Derived Crashworthiness Ratings." Proceedings (*2011) Australasian Road Safety Research Policing Education Conference,* November, Perth, Australia.

6. Assaad, Zena, and Christine Boshuijzen-van Burken. “Ethics and Safety of Human-Machine Teaming,” 1–8. Edinburgh, United Kingdom, 2023.

7. Lee, J. D, and K. A See. “Trust in Automation: Designing for Appropriate Reliance.” *Human Factors* 46, no. 1 (2004): 50–80.

8. Ullrich, Daniel, Andreas Butz, and Sarah Diefenbach. “The Development of Overtrust: An Empirical Simulation and Psychological Analysis in the Context of Human–Robot Interaction.” *Frontiers in Robotics and AI* 8 (2021).