Trustworthy AI: An operator's view (but a systems-engineering approach)

Ben Keith explores the topic of AI assurance, and understanding the balancing act of trusting AI systems.

FRAZER-NASH CONSULTANCY
A KBR COMPANY

A key tenet of successful AI is that of assurance. Understanding when and where to trust AI decisions is key to reaping the benefits it has to offer.

The art of being a professional mariner is about making decisions, whether related to shipping situations; routing and weather; engine configurations, through to threat evaluation and weapon allocation in a military context.

The 'OODA loop' (Observe, Orient, Decision, Action) decision-making cycle has been in use by the military for decades. In a human context, assurance of the accuracy of each phase can be achieved through techniques such as observing the process and challenging assumptions. However, if part of or all of the system-of-systems that delivers the action is automated, assurance of the process is difficult, particularly if the OODA loop operates at machine speed.

This article takes an operator's view of the challenges, with a brief introduction of the technical: It is proposed that assurance of the Observe, Orient, and Decide functions in a system-of-systems involving AI is vital, and can be achieved by balancing an objective measure of system trustworthiness with a subjective assessment of how much users trust the system.

Decision Making – a human context

The 'OODA loop' (Observe, Orient, Decision, Action) is a widely-known decision-making model that originated in the US Air Force but is now used in a variety of civilian and military fields. It translates well from human to machine, with automation bringing benefits of assimilation of vast quantities of data, rapid assessment, and machine-speed repeatability. However, operating at machine speed, governed by AI with impenetrable algorithms and overlaid with machine learning, all stages of the decision-making cycle become opaque, with only the 'action' being observable.

To place this into context, the OODA loop output in a human based, 'Rules of the Road' scenario is neatly demonstrated in the well-practiced phrase of a Royal Navy (RN) Officer of the Watch (OOW) in a 'shipping report':

> *'Captain Sir/Ma'am – at red 40, range 1 mile, I have a vessel engaged in fishing* **[Observation]**
>
> *we are 30 degrees on their port bow and its bearing is drawing left*
>
> *with a Closest Point of Approach of 2 cables to port.* **[Orientation]**
>
> *It is a Rule 18 situation, and we are the give way vessel.* **[Decision]**
>
> *My intention is to alter course 10 degrees to starboard to open the CPA to 5c, it is safe to do so.* **[Action]**

Assurance of that decision-making cycle starts at basic training, and is formalised through qualifications, such as STCW (International Convention for Standards of Training, Certification and Watchkeeping for seafarers). It is then built upon by supervised experience, with defined parameters that the OOW can act within. Assurance that they have oriented themselves based on correct interpretation of all the information would come from questioning the decision and verifying key pieces of information – for example, an oft repeated phrase on a bridge of a RN ship is 'don't trust the radar – look out of the window'!

There are many softer skills and human factors that also contribute to decision-making assurance, such as, how much sleep has the OOW had; what is their

arousal levels; are they under any other stress; how confident are they – does their tone of voice give away anything; have they made mistakes before. Most mariners with experience of Command have a story about being jolted awake in the middle of the night by something that 'didn't feel quite right', often called 'intuition', but more accurately should be called 'experience' – with a rapid re-location to the bridge preventing a more serious incident developing.

It could be assumed that getting to the right 'Action' is the point of the OODA loop, which would imply that the journey to get there is immaterial. That would be the case if each incident is taken in isolation, but, for a OOW to improve and gain experiential learning ('machine learning' in an automated system) throughout their career, the component parts of the OODA loop are just as important. An action which is apparently correct, but based on an incorrect interpretation of the situation is more insidious than an obviously incorrect decision, which can be observed and corrected.

Decision making at machine speed

Substitute AI or 'autonomy' into a maritime system-of-systems, particularly in autonomous navigation and collision avoidance, and it is virtually impossible to assure each part of the OODA loop: The decision/action which is the output of the OODA loop can be assured through a variety of means such as repeated testing in live or synthetic environments. Alternatively, if required assurance levels cannot be met, risk can be managed through imposing control such as limiting operating areas to account for its capabilities or requiring a human to be embarked for supervision.

But how can we assure the 'Observe' and 'Orient' functions? How do we know why a decision has been made, and therefore assure its repeatability? This assurance is important as the level of automation in a system increases; particularly important if AI is applied; and vital if machine learning is overlaid.

A vignette from early autonomy trials of a RN vessel can illustrate the necessity of understanding the entirety of the OODA loop: Whilst transiting an empty Solent on a clear day at speed, the autonomous vessel (jet propelled) came to an instant stop (causing consternation to the human inhabitants!) just before the Coastguard helicopter flew past the bow at 100kts / 100 feet. The action of giving way to an unknown, bow-crossing AIS fitted vessel doing 100kts was probably correct, but it relied on interpretation of an AIS fitted vehicle doing 100kts being on the surface, vice the air.

Trust in automated systems

The key to assuring the safety of an AI enabled system-of-systems is to build trust in the system, and achieve 'calibrated trust', which is a balance of:

- User trust – a subjective assessment of how much you trust the system, and;
- System trustworthiness – an objective measure of how much can the system be trusted?

If it is not balanced – there is either too much trust in the system, leading to 'overtrust', or a position where the supervisor 'knows best', in a position of 'undertrust'.

History is littered with accidents and incidents relating to both of those situations – 'overtrust' resulting in misguided, incorrect or damaging actions taken by the system with no human input or supervision when there should be; 'undertrust' resulting in 'human knows best' and taking contrary, alternative or abortive action by the human based on flawed assessment of the situation.

'Calibrated trust' as a concept can apply to all system of systems, including human centred ones; consider the vignette of a RIB recovery to a RN ship in a navigationally-constrained situation in which the author was involved. A situation of 'calibrated trust' with a well-briefed plan and experienced team quickly escalated to become an incident when an inexperienced coxswain was substituted into the 'system'. This moved the system to be out of balance, as the system could not deliver what was being asked of it – an 'overtrust' situation, putting the entire ship into danger.

Finding Calibrated Trust

So how to find 'calibrated trust'? Frazer-Nash and University of Bristol have used a systems-engineering approach to design a set of 'trustworthiness categories', which can be prioritised for a system based on the Concept of Operations and risk analysis. These can be assessed against known/published standards to define a Trust Quality Assessment with performance monitoring and a vital feedback loop.

Work is ongoing in the UK, US, and EU to define the set of AI standards and laws against which to benchmark AI trustworthiness, and Frazer-Nash is working with areas of the military, such as Defence Science and Technology Laboratory (Dstl) to assess applicability in a military context, such as Automatic Target Recognition, fire control and flight control systems.

It is still early days, with technological advances outpacing our ability to regulate and enable. However, there is opportunity for the UK and in-particular the UK maritime industry, to be at the forefront of designing and adopting AI trust standards; advancing technical capability and importantly enabling a suitable regulatory regime that maximises innovation opportunity for the UK but constrains and protects against nefarious use.