The Test and Evaluation of Intelligent Autonomous Systems (IAS)

H Figueiredo CEng, MIET, D Cook CEng, MIET, W Biggs CEng MIET

[a]*QinetiQ Ltd UK*

SYNOPSIS

Intelligent autonomous systems (IAS) are set to become a feature of future defence programmes, and their introduction will pose challenges to traditional systems engineering and acquisition practice. Whilst the need for operational and technical assurance will endure, the need to manage programmes that deliver iteratively and which are continually evolving requires fresh thinking. New increments may well be undergoing acceptance testing against a backdrop of continual developments to higher level concepts of operation – there may be no stable baseline. Furthermore, the unbounded and potentially non-deterministic nature of IAS means testing alone is unlikely to provide satisfactory assurance, especially for systems that are able to learn from previous mission data. Lastly, at the very core of what is referred to as human-autonomy teaming is a notion of trust by the operator of the IAS, a trust which builds through development, integration, training and deployment, implying a blurring of boundaries between that which is technical assurance and that which is operational. In response to these challenges, QinetiQ are investing in UK test and evaluation infrastructure and developing, with partners, approaches that will mitigate the risks posed by these new technologies. As outlined in this paper, these include distributed live, virtual and constructive facilities, brokered policy enforcement by software and developing a body of trusted software components. The paper argues how these developments address the identified challenges, highlighting remaining gaps and drawing on evidence from ongoing UK MOD research and development; it concludes with the approaching investment and programmatic choices that will need to be made to ensure best-for-enterprise outcomes.

Keywords: Maritime; Autonomy; Test and Evaluation; Artificial Intelligence.

1 Introduction

The UK MOD Defence Industrial Strategy (DIS)[1] points to the importance of Test and Evaluation (T&E) recognising that *"A UK-based Integrated Test and Evaluation capability is essential for quality assurance and some safety and operational security needs. We do not consider it essential to carry out the testing onshore, yet it is vital that we retain the capability to understand, interpret and direct the testing to meet our performance and safety standards."* There is therefore an enduring UK requirement to ensure that defence systems meet required specifications and that the risk to life is clearly understood and can be appropriately mitigated.

Intelligent Autonomous Systems (IAS) are one of several emerging technologies under the banner of the Fourth Industrial Revolution (4iR) that are being applied across a range of environments as part of a widespread shift in the technology we use every day. These systems build on technology areas such as Artificial Intelligence and Machine Learning where technology seeks to optimise and improve performance and behaviour over time, learning and adjusting as new data becomes available. The potential for these systems to deliver an operational advantage in defence, security and critical national infrastructure is huge. But that potential will only be realised if these systems have been evaluated for safe and practical deployment in such high-criticality environments.

The problem is that, by their very nature, these systems exhibit no stable end state against which they can be tested using traditional engineering methods. These approaches will only provide a snapshot, and repeating the test for an adaptive system will potentially give a different result every time as the system may have learned during the tests to which it has been subjected. Herein lies the conundrum – how do you evaluate whether a system is fit for deployment when the system will change its behaviour in response to the testing itself? Even those IAS which are not adaptive and which instead apply deterministic approaches can be problematic to traditional state based assurance[2], so new approaches are required. A further factor is rapidly changing concepts of employment (CONEMP) and requirements, driven by the disruptive nature of IAS - as a result there may be no stable baseline.

Despite progress in recent years with greater use of model based approaches, the challenges in delivering comprehensive T&E of IAS still remain and can be identified through-out the lifecycle[3]; for example:

[1] Defence Industrial Strategy, Defence White Paper, Secretary of State for Defence, Dec. 2005

[2] For example a simple vision based obstacle avoidance system can quickly generate an almost infinite set of possibilities for review.

Authors' Biographies

Derek Cook is Principal Systems Engineer with over 35 years' experience of command and control systems and instrumentation integration in a safety critical environment, which includes the recent integration of the live Range environment with virtual and constructive synthetic environments.

Hector Figueiredo is a autonomy capability with over 30 years' experience in Engineering of Aircraft and Intelligent Systems.

Bill Biggs leads QinetiQ's work on Autonomy including the QinetiQ Maritime Autonomy Centre, with its particular focus on unmanned systems in the maritime environment.

[3] Defence typically employs a CADMID approach, but the same considerations apply to iterative or spiral methods.

 http://doi.org/10.24868/issn.2515-8171.2019.001

- At the requirements stage, it is challenging to capture user requirements that translate consistently to systems requirements and operational T&E when the operational requirements are changing; this is a challenge for all complex systems but the inclusion of AI is an exacerbating factor.
- At the design and development stage, changing requirements creates a challenge to gather and record relevant evidence at simulation/ run-time to support testing. Models need to be validated and support cumulative run-time testing and evidence gathering. However, in addition to ensuring software integrity regarding the models, there are additional challenges regarding trust in the data required to validate the models given that training data for AI algorithms are labelled by humans which may be prone to error.
- At the operational testing stage, there is a need to inspire user confidence - defining trust metrics that provide confidence in the system to be deployed in changing uncertain environments; that they will perform reliably – being able to extrapolate from known conditions.

T&E also provides users with confidence that the IAS will be robust and/or behave consistently to real world sensor inputs and changing environmental conditions and that as these degrade, impact on mission plans and delivery of effect is clearly understood. User confidence and trust is key in tasking IAS, but it is also fundamental to more collaborative and potentially pre-cursor approaches, such as Manned Unmanned Teaming (MUMT) or Human Autonomy Teaming (HAT). This white paper presents QinetiQ's vision for how the federation of Test and Evaluation capabilities with Customer research, experimentation and development capabilities underpinned by Live, Virtual & Constructive capability, allied with a range of new techniques can support the T&E of IAS to meet the identified challenges.

2 Live, Virtual & Constructive (LVC) - An Evolutionary Approach

Recognising real world constraints of time and cost, adopting a risk balanced, evidence based approach to testing is key to the delivery of affordable and assured IAS programmes. Increasingly this is enabled by Live/Virtual/Constructive (LVC) capabilities, exploring operations within a mission context as part of an agile process based around a short-term development–test–development cycle. Some key benefits of LVC with respect to supporting T&E are:

- An environment that enables models/capability/technology to be rapidly developed and tested
- Early exploration of concepts to help to define long-term requirements and development strategy
- Human-in-the-loop use cases to explore safely and identify trust issues
- Low cost, low maturity testing prior to more costly methods as systems and technologies mature
- Iterative integration of multiple systems

Figure 1: Images captured during trials using the ARENA Synthetic Environment & the SUAS available for Live Virtual & Constructive trials

QinetiQ have increasingly been applying LVC techniques to their autonomy research and development. Figure 1 shows QinetiQ's autonomy LVC capability which consists of its Autonomy Research Environment for Novel Architectures (ARENA)[4] and Small Unmanned Aircraft System (SUAS) Autonomy Research Capability. The SUAS capability can in principle be swapped out for other unmanned systems evidenced by the recent transition of this capability in support of unmanned ground systems for Autonomous Last Mile Re-Supply (ALMRS)[5]. Similarly these capabilities have increasingly been worked into other domains, notably the maritime, with both ARENA and SUAS being employed as part of the MAPLE 4 programme[6].

[4] N. Swain et al, Adaptable Autonomy Rapid Prototyping Assessment, Aug. 2016.

[5] W. Kennedy-Scott et al, TITAN robot as an open and modular ROS platform – ALMRS Final Report, March 2018

[6] P Smith et al, 'Securing interoperable and integrated command and control of unmanned systems – validating the UK MAPLE architecture', EAAW 2019

 http://doi.org/10.24868/issn.2515-8171.2019.001

Ideally, an LVC capability would be capable of being integrated as part of a larger network with other, potentially more immersive/representative simulations with remote sites where live elements of LVC trials are undertaken, enabling early T&E and helping to achieve technology transition and exploitation. This is the intent for ARENA[7], particularly in the air domain, with a standing facility at Farnborough integrated with the West Wales UAV Centre and the Trials Control System at Aberporth and Hebrides Ranges and, subject to agreement, the Air Battlespace Training Centre (ABTC) at RAF Waddington, leveraging investment to enable the Long-Term Partnering Agreement (LPTA) to meet its original intent servicing the needs of future systems. A key capability of test or evaluation systems is the ability to log data from components and nearly all of the system and subsystem components within the ARENA LVC environment can produce logs for later review, a feature likely to be increasingly important as training data for machine learning is required.

ARENA includes components that enable connection via Data Distribution Service (DDS) messaging; the DDS/Datalink library allows interface between any/all real-world datalink formats/protocols including Link 16, the versatile message format (VMF), Link 22 etc. These are used with the ARENA LVC to send situation awareness information and tasking information between systems, particularly higher technology readiness level (TRL) components.

Experience gained on research and development and technology demonstrator programmes has shown that LVC through the ARENA/SUAS pairing is a key enabler to gain confidence in the safety, sensitivity to real world inputs and concepts of use in a cost effective and risk balanced way. It provides an environment to build and test models of system elements; integrate iteratively developed sub-systems and a building block environment to collect a body of evidence as systems and capability matures. This LVC approach featured extensively in QinetiQ work for Dstl into HAT and MUMT. In addition to developing portable control interfaces, the application of human centred metrics has allowed exploration of operator trust in IAS, a trust which builds through development, integration, training and deployment (customer trials), implying a blurring of boundaries between that which is technical assurance and that which is operational.

3 Key Enablers – Facilities & Infrastructure

In recent years, QinetiQ have initiated a strategic drive towards the federation of research and test facilities to allow more effective collaboration between programme partners to drive pace into development programmes, to enable earlier Test and Evaluation and minimise the need for deployment teams to be located away from their home sites for extended durations and in large numbers.

3.1 Instrumented Test Range Overview

QinetiQ operates a number of Instrumented Ranges for Test, Evaluation and Training on behalf of the UK Ministry of Defence (MOD) under the LTPA. The Test Ranges at Aberporth and Hebrides for example, are fully instrumented Danger Areas where the test and evaluation of complex systems can be safely undertaken. Test and Evaluation Trials are designed not only to meet the Customer's mission objectives, but to also keep both participants and non-participants safe from hazardous activities.

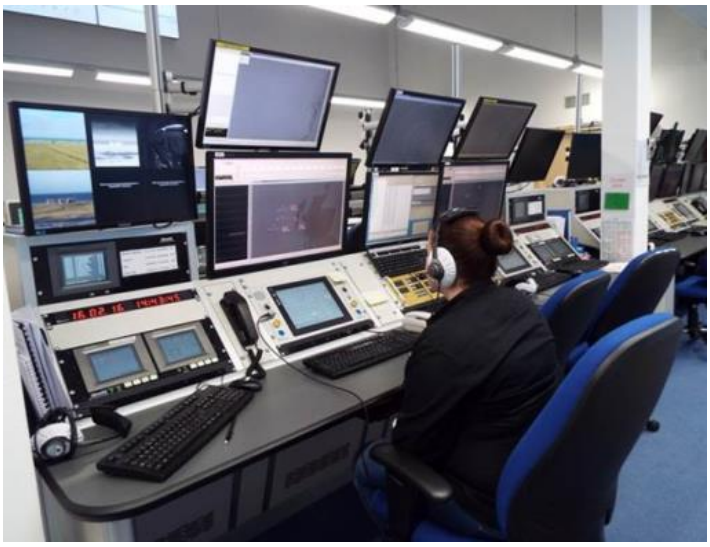

Figure 2; Aberporth Range Control

[7] Already delivered at low technology readiness level as part of the Modular Air Command and Control System project in 2017

 http://doi.org/10.24868/issn.2515-8171.2019.001

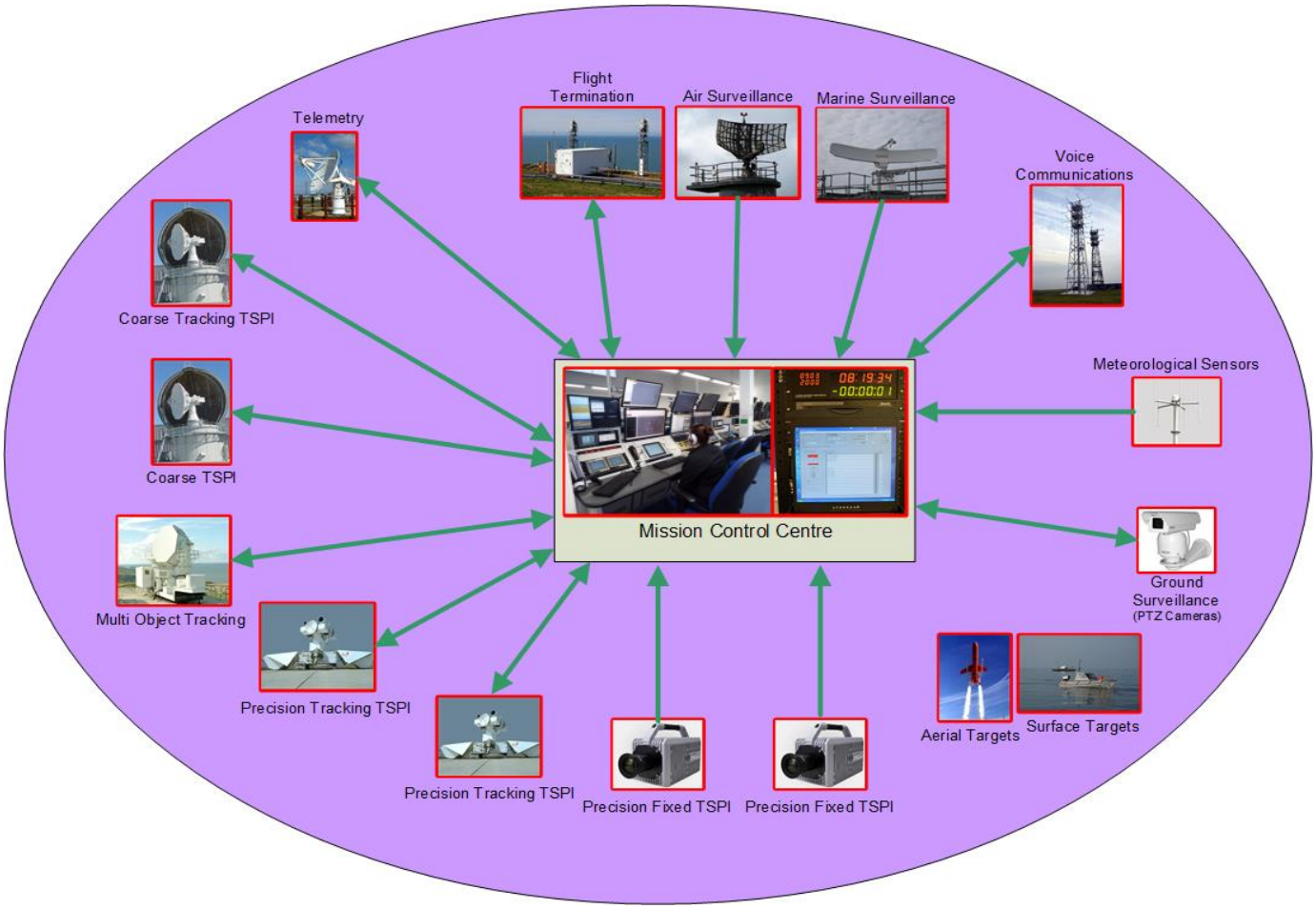

Figure 3; Typical capabilities available at a QinetiQ test range

The Hebrides range provides the largest instrumented Test Range in Europe, suitable for the test and evaluation of long range systems. The Aberporth Test Range is smaller in size, but is likely to be more suited to the initial T&E of IAS due to the proximity of facilities such as the Snowdonia Aerospace Centre, which could be suitable for the take off and recovery of armed IAS platforms, along with the fact that the shorter range activities that it can support can be more easily covered by optical instrumentation to independently capture the behavioural data of new systems. Both the Hebrides and Aberporth have supported, and are increasingly utilised for both maritime and air unmanned systems trials work. A very simplified block diagram of the typical capabilities of a Test Range is shown in Figure 3.

4 Instrumented Air Test Range LVC Integration

Under internally funded research and development, QinetiQ has successfully integrated its Trials Control System with its simulation integration toolkit, AIME (Architecture Independent Modelling Environment). AIME is used to provide a bridge between trials control system (TCS) network protocols and the Distributed Interactive Simulation (DIS) Protocol to connect to representative simulation capabilities. For the purposes of the LVC IRAD work, the scope of integration with simulation capability has been the Training and Innovation Facility (TIF) at QinetiQ Farnborough. However, it could be any DIS protocol compliant simulation capability anywhere in the world if there is a suitable network connection, or AIME can be used for protocol conversion if required. The result of this integration work was that all the key actors, including range staff, could collaborate much more effectively on the design and derisking of complex trials with significant safety management challenges.

The T&E IRAD programme initially demonstrated the concept between the Aberporth Range and the Farnborough TIF in 2017. A more mature instantiation between the Hebrides Range and Farnborough has recently been completed and was demonstrated in March 2019. QinetiQ's architecture vision is for a "hub and spoke" model where Farnborough is facilitating a "Capability Generation and Assurance Hub", making the connections between the remote Ranges to a more centrally accessible location, from where it will be easier to federate with other capabilities as needed. The concept of the Ranges live and virtual integration is shown in Figure 4 below.

 http://doi.org/10.24868/issn.2515-8171.2019.001

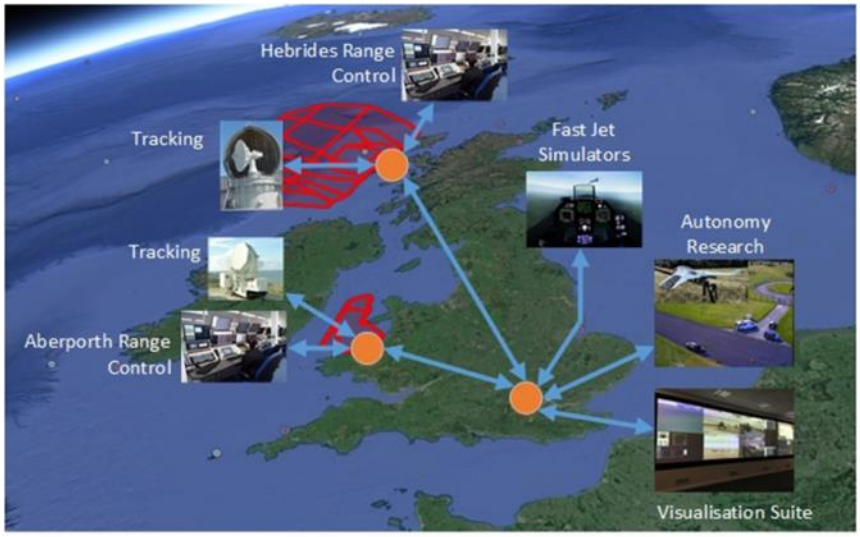

Figure 4; Integration of Test Ranges with Simulation Environments

The simulation and current autonomy capabilities are represented by QinetiQ facilities, however the concept of the hub is to provide a portal for making secure connections to other research, experimentation, simulation and test and evaluation facilities that needed to be federated, and, subject to approved connection, could be located anywhere. As such this hub and spoke model provides a scalable, open architecture and a design pattern for connecting to other capabilities. These capabilities could include additional MOD capabilities or the capabilities provided by other organisations and companies.

As can be seen above, the ranges are integrated to the hub at QinetiQ Farnborough, which is more readily accessible than the remote capabilities (the Hebrides Range is approximately 900km from Farnborough). Via the LVC integration capabilities, the remote ranges can be connected to DIS compliant simulation capabilities (other protocols could be supported), such as hi-fidelity, representative fast jet simulators or experimentation and research and development rigs for autonomy research. A similar approach to integrate with the shore integration facilities at Portsdown Technology Park is under consideration as a follow on development.

LVC integration to Customer trials has been successfully achieved several times, where the live trials data is exported to the Synthetic Environment, which allows the SE to make use of live real time "pattern of life data". To enable this work to take place, formal safety analysis of the LVC exercises have been undertaken and the necessary security approvals have been obtained to make the connections. To date the approved connections have been live to synthetic. QinetiQ considers that any requirement to go from synthetic would initially need to be on a case by case basis.

5 IAS Test and Evaluation Use Case

Having briefly described the Test Range Capabilities and how the LVC is integrated, this section now considers the use case of the Test and Evaluation of a new IAS, using manned unmanned teaming with fast aircraft as an exemplar, and the evolutionary steps that could form part of an overall IAS T&E programme. The key challenges for IAS T&E is to derive a program that ensures that adaptive algorithms which are learning from data sets can be verified to the appropriate safety assurance level, when they are inherently non-deterministic. There has been a lot of work[8][9][10] to develop methods that provide performance guarantees based on model-checking and formal methods. The models must fully describe the autonomy performance and exhaustively tested for exceptions that break the specifications. Perceived drawbacks are that the models must first fully describe the autonomy, and that test engineers must have full access to the models. As autonomous systems increase in complexity and the proprietary nature of software limits access, these limitations raise questions regarding trust and assurance in the test results. Answering these questions requires new approaches and these are explored in more detail in section 6, but none of these remove the need for progressive T&E and user involvement, moving from model based work, to synthetic to LVC activity. Indeed, these new components and methods are dependent on such a progressive T&E programme. Delivering such a progressive programme, requires a scenario driven methodology with a consideration of the mission profiles that will be required, expressed in a manner that reflects that it may not be possible to fully anticipate all outcomes. The exploration of scenario outcomes also needs to consider failed outcomes as well as successful outcomes and conditions that are not part of normal operating, for example, extreme weather or threat of collision with other vehicles.

This scenario driven consideration can then be used to construct a system assurance programme which will generate

[8] Formal methods for learning and reconfiguration in autonomous systems, T. Wilkenson & M. Butler, University of Southampton, 2015

[9] Formal Methods, 22nd International Symposium, FM 2018

[10] 14 International workshop on Advances in model based testing, (A-MOST), 2018

the data sets required to stimulate the IAS, and gather and assess the required evidence to the required level of confidence. In all cases the assumption is that a precursor activity is model based design and testing. A high level outline of possible evolutionary steps in a T&E is now presented, with the logical sequence broadly being:

- Step 1 - Synthetic Constructs only
- Step 2 - LVC – Live cooperating platforms into a Synthetic Environment
- Step 3 - LVC – Live IAS under test, with cooperating platforms in a Synthetic Environment
- Step 4 - LVC – Live IAS and cooperating platforms in segregated live environments, feeding a Synthetic Environment model (or Digital Twin) of the IAS to aid in model validation
- Step 5 (and beyond) - Further variations on Step 4, as confidence is built to reduce the segregation limits between the platforms

Step 1 – Synthetic Constructs only
In Step 1, there would be no live activity and no need for LVC integration, but live environments and assets would be replicated in the synthetic world as the first step in scenario planning, de-risking, rehearsal and execution.
Live data sets from previous trials (where approval can be obtained from the Customer for IAS T&E use) might be used as part of the data sets. Under the LTPA, QinetiQ are custodians of Customer trials data captured by Range instrumentation. The records are held for at least seven years.

Step 2 – LVC – Live cooperating platforms into a Synthetic Environment (Figure 5)
In this step, there is no live IAS; live data from various other platforms is used to stimulate IAS algorithms on their development rigs during initial development and software module testing. The aim would be to get the IAS algorithms proven to an agreed level of confidence in controlled conditions and in response to the scenarios identified in the T&E programme. The same test cases used in Step 1 may be repeated, but now synthetic data sets are replaced with live data sets with all of the subtleties and variations that live data will contain.

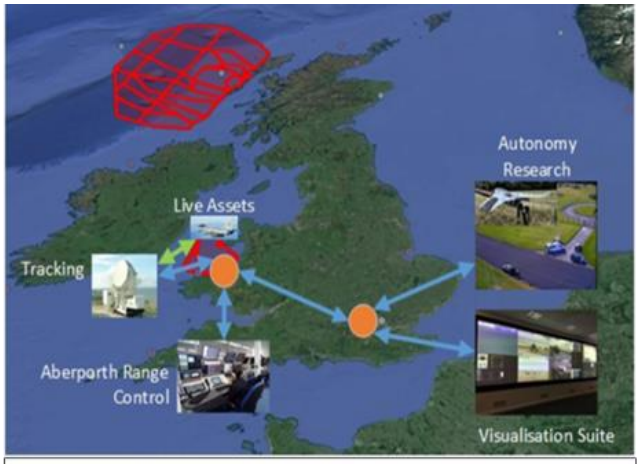

Figure 5; Step 2 – LVC – Live cooperating platforms into a Synthetic Environment

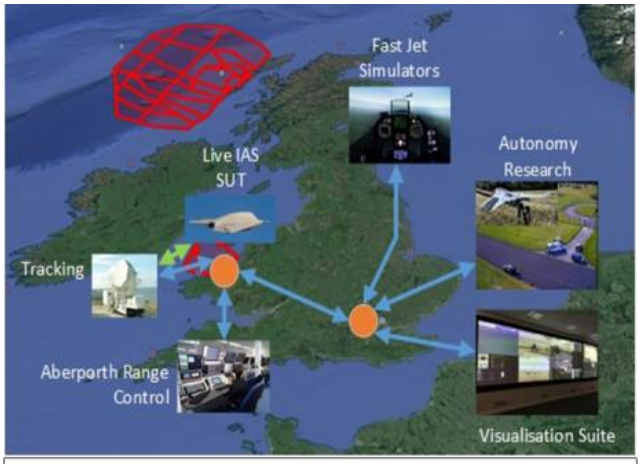

Figure 6; Step 3 – LVC – Live IAS under test, with cooperating platforms in a Synthetic Environment

The scenario data sets would be captured so they can be reused, for example during regression testing as problems in the IAS algorithms are found and fixed.

Step 3 – LVC – Live IAS under test, with cooperating platforms in a Synthetic Environment (Figure 6)
In this step, the IAS System Under Test (SUT) is now mature enough to fly in the live environment as part of the T&E programme, however it has no approval to fly in the same airspace as other manned platforms. In a collaborative autonomy scenario, there may be series of sub-steps ranging from a single IAS SUT to a swarm of them. Live manned platforms are represented in the LVC construct as simulated entities. For collaborative manned/unmanned teaming scenarios, the pilots in the fast jet simulators are flying in a virtual representation of the Range and under the direction of Range staff as if they were live entities, with their synthetic data being fed to the IAS as if it were live. In this scenario live data could still be fed to the research and development rigs and models running on them as an aid to model validation. For example a step to IAS swarms may have a single live IAS, communicating with and coordinating a virtual IAS swarm.

Step 4 – LVC – Live IAS and cooperating platforms in segregated live environments
Once the preceding steps have provided the right level of confidence it is time to undertake live flying activity of both cooperating platforms and the IAS. At this point care is still taken to ensure safe segregation. In the first instance this may require the IAS to fly in different segregated air space. Any rogue IAS activity will result in flight termination. The operating areas are spaced such that the Danger Area associated with the weapon (which is fixed to the IAS until the point of firing)

does not encroach the Fast Jet operating area when the IAS SUT remains within its permitted operating area. If the IAS is placing the fast jet at risk then the mission will be terminated and the Fast Jet will break off. The IAS flight would be terminated if it is not under positive control.

Step 5 – Further variations on Step 4
The same concept as described in Step 4 would apply. In this Step as the T&E assurance programme builds confidence in the integrity of the IAS, then separation distances are slowly reduced. Profiles are designed to minimise risk to the live platforms. For example, live platforms need to stay outside of weapon danger zones. Beyond Step 5 lies user trials and training.

6 Challenges of Artificial Intelligence (AI) and T&E

As set out earlier, the desire to exploit more Artificial Intelligence and Machine Learning where technology seeks to optimise and improve performance and behaviour over time, learning and adjusting as new data becomes available demands new approaches to manage adaptivity. Within ARENA, research has been undertaken in to developing the COMPACT[11] (Configurable Operating Model Policy Automation Control of Tasks) architecture for control of IAS; COMPACT selects techniques and operating constraints based on a rule-based trusted reasoning process applying a set of pre-determined (and configurable) rules. As such it can monitor and control IAS using variable levels of autonomy and automation. The rules dictate behaviour appropriate to mission context including class of airspace, Rules of Engagement (RoE) and mission phase. These behaviours are not necessarily "intelligent" and in some circumstances predictability will be more important than optimising. Critically COMPACT itself is fully deterministic and therefore testable, verifiable, certifiable. As a policy guard it can be accredited, potentially removing the need for the autonomy algorithms themselves to be subjected to the same level of rigorous test. The rules invoke system functions with appropriate technology/techniques and these functions may be intelligent or procedural. So COMPACT, with the caveat that it is a research product and is not yet a fielded or complete and proven capability, enables accreditation, and existing assurance requirements to be met. But on its own it will not be sufficient as an understanding of system performance and decision making and user trust requires a range of other approaches to also be applied. These include:

1. In addition to LVC, consideration should be given to the qualification of adaptable intelligent systems through the lens of human factors and behavioural science in collaboration with traditional engineering practices. The feasibility of adapting the techniques we use to for establishing human performance and behaviour should be considered so they can be employed for the evaluation of technology. There is already work underway in this area – including studies in the US investigating the use of virtual mazes to test the behavioural psychology and cognitive skills of artificial intelligence systems, and papers from research teams in China on the adaptation of common psychometric and IQ tests to assess their abilities, attitudes and knowledge traits.
2. Consideration of the through-life requirements for maintaining these systems via regular re-qualification, robust configuration management and the adoption of appropriate regulation and legislation.
3. The generation of certifiable training data and the challenges of sanitising and labelling the data for the use of machine learning and AI algorithms.
4. Consideration of the professional accreditation of intelligent systems designers to ensure they understand the unique testing requirements for assurance in defence and security environments and how to inject these test and evaluation criteria early on in the development pathway.

7 Conclusions

This paper has considered at a high level, the challenges faced by the T&E of new IAS. It considers how a Live/Virtual/Constructive approach with Live integration between the UK MOD Ranges, integrated to federated research and experimentation test beds and synthetic environments can provide a structured and methodical means for progressively moving through the T&E Lifecycle and gathering the required assurance evidence necessary to provide the right level of confidence in the IAS under test. Such an approach can also gather large amounts of assured data, useful in the development of machine learning systems. Whilst the examples given have primarily focused on air systems, the approach and toolset are equally valid in the maritime environment and with maritime facilities and this is an area of development as QinetiQ seek to broaden and deepen their capability. Recognising the particular challenges or learning and adaptive systems, an accreditable policy guard, such as COMPACT will increasingly be important in progressing and maturing IAS capabilities. Further development of an application such as COMPACT is therefore a priority if defence fielding of IAS is to be achieved.

Beyond development of the LVC framework and COMPACT, the authors conclude that further work is required and should be progressed into the qualification of IAS systems through the lens of human factors and behavioural science in collaboration with traditional engineering practices and that more priority is given to the collection of certifiable and labelled training data from ongoing trials and development activity. Lastly, full assurance suggests a need for some form of professional accreditation of intelligent systems designers to ensure they understand the unique testing requirements for assurance in defence, including the implications of regulation and legislation.

[11] RJ. Cottrell, MJ. Thomas, Hybrid Architectures for Adaptable Autonomy V1.3, Nov 2013

 http://doi.org/10.24868/issn.2515-8171.2019.001