Measuring Effectiveness of Human Autonomy Teaming

Helen Lashley, Aaron Thorpe, Robert Taylor and Antony Grabham
Defence Science and Technology Laboratory
Dstl Portsdown West, Fareham, PO17 6AD
UNITED KINGDOM

hjkeirl@dstl.gov.uk

ABSTRACT

The aim of this work is to test and evaluate the effectiveness of HAT, specifically the interactions and dialogue between a human operator and autonomy system in joint, collaborative manned-unmanned operations.

Throughout the comprehensive UK MoD Dstl programme of R&D on Autonomy and Mission Systems, a number of different metrics have been developed predominantly focused on exploiting the critical decision methodology to address the different constituent parts of the complete HAT System.

The Human task work and teamwork elements of the HAT system, through CAPTEAM which is designed to estimate mission efficiency by the metrication of Reward and Effort associated with critical mission events and decision processes. The Autonomy Task Work component of the HAT system is addressed through a broken down multi-dimensional Trustworthiness scale. The entirety of the HAT system can be assessed through a combination of the HAT Capability Maturity Model and the Risk Assessment. The new component of the UK HAT System assessment methodologies is the REMEDE assessment protocol. The protocol itself is focused on the Human Autonomy Teamwork component of the entire system, and has been developed from a variety of information processing, communication and team work models.

At this stage, the REMEDE protocol has been selected for use with the STRATUS project under the UK MoD Dstl Autonomy Research programme, with a specific instantiation developed for the specifics of the trial. This trial will serve as the initial opportunity to undertake verification and validation testing of the REMEDE protocol, the outcomes of which will allow a greater understanding of the discrimination and sensitivity of the data captured.

1.0 INTRODUCTION

Human-Autonomy Teaming (HAT) - the creation of a mutually supportive, co-operative, collaborative partnership between the human user and advanced “intelligent” automation technology - has been recognised as the necessary strategic, human-system integration design objective for Unmanned Systems (UxS). HAT is aimed at delivering efficiently and effectively the required agile, adaptive, context sensitive, multi-mission capability, whilst enabling meaningful and effective Command and Control (C2).

Prior to the advent of Artificial Intelligence (AI) and before the introduction of decision making technology, the human commander/operator retained sole responsibility for providing cognitive decision making capability. The commander/operator uses sensing, thinking, values, reasoning, learning, knowledge and memory, coupled with judgement, vision and imagination, shared and developed through social communication and dialogue. Cognitive capability enables human judgement to be reasonable, dependable and reliable in dealing with uncertainties and ambiguity. But it also enables humans to be innovative and creative, thereby anticipating options, contingencies, risks, opportunities and threats. This creative cognitive

capability provides the essential context sensitivity, flexibility, agility and adaptability needed for decision superiority in the complex, dynamic military environment.

Concurrently, the provision of effective test and evaluation methodology for proving HAT efficacy, with verification and validation of human-system performance and effectiveness, has been identified as presenting a significant technical challenge as described by R Taylor [1], [2].

The aim of this work is to describe the test and evaluation methodologies used in determining the effectiveness of HAT, specifically the interactions and dialogue between a human operator and autonomy system in joint, collaborative manned-unmanned operations.

2.0 MILITARY CONTEXT (RATIONALE)

Remote and automated systems employing unmanned and remotely operated platforms, and systems employing embedded automation and self-governing autonomous functionality, are being increasingly employed as enablers of military capability. However, it is widely recognised that in the complex, dynamic and uncertain environment of military operations, human effectiveness and HAT is needed, as an axiomatic design imperative, for autonomous systems to be assuredly safe and effective.

Test methods are needed that are designed to achieve the levels of proof required to support evidence-based decision making in UK MoD Defence Equipment procurement. Selection of test methods involves balancing requirements for innovation and demonstration (e.g. use case testing), compared with the need for robust performance-based trade-space analysis, valuing discrimination power and optimisation (e.g. multi-variate performance testing). Innovation test methods can be less rigorous experimental design with changes to technology, adaptations to Tactics, Techniques, Plans and Procedures (TTPs) and learning encouraged throughout the duration of a trial. The data and outcomes are used as indicators for future research and lessons identified rather than statistically significant results. Discrimination test methods are more robust and rigorous with dependent and independent variables. The trials outputs are more defined and lessons and conclusions can be drawn directly from the data. Figure 1 illustrates the distinction between innovation and discrimination test methods, (R Taylor 2015 [1]). A balanced approach using both test methods has been employed by UK MoD for HAT research, resulting in quantitative and qualitative data as well as validated assessment methodologies for testing HAT effectiveness, as described in this paper.

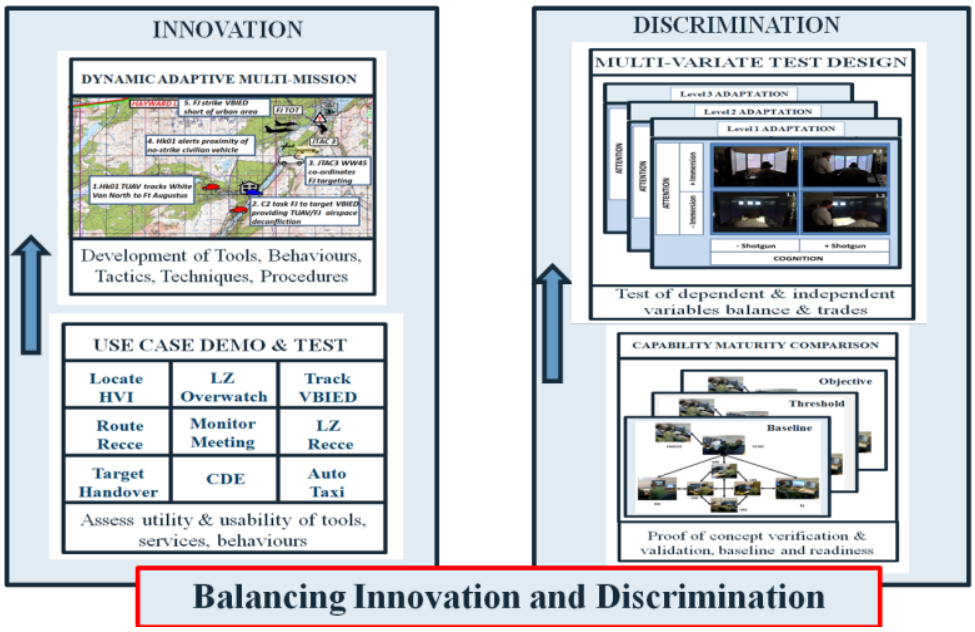

Figure 1: Applicable Test Methods

2.1 Critical Decision Methodology

Underpinning the test methodology, and providing the required human-effectiveness centric focus, Dstl has developed a cognitive capability testing approach, based on dynamic Mission Critical Decision Making (MCDM) methodology. MCDM focuses the locus of testing on Course of Action (CoA) adjustment responses to dynamic mission critical events. Uncertainties in the military combat environment affect understanding of risks and opportunities associated with delivering the command intent and in the performance of associated tasks, with dynamic targets and threats, resulting in adjustments and changes to CoA tactics, techniques, plans, procedures, all affecting the delivery of effects. (Taylor and Grabham, 2012 [3])

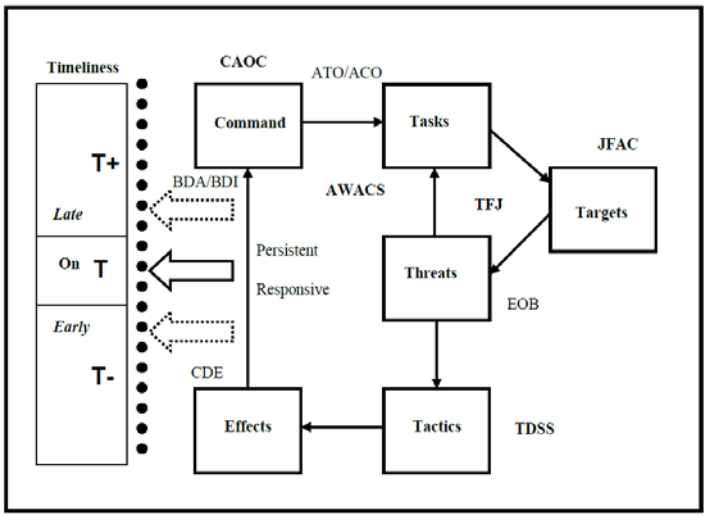

Figure 2: Model of Dynamic Air Mission Management

MCDM employs objective, task analysis and effects-based Measure of Effectiveness and Measures of Performance (MoE/MoP), coupled with subjective decision-based metrics for estimation of human and mission effectiveness (Collaborative Adaptability Proficiency Test Evaluation Assessment Methodology (CAPTEAM)). Essentially, the aim is to capture the agility (speed) and adaptability (quality) of mission commander/operator decision-making in dynamic mission timeline context as illustrated in Figure 3.

Subsequently, under the UK Autonomy and Mission Systems Research Programme, the MCDM methodology was further developed to measure collaborative adaptability proficiency with autonomous systems concepts and technologies. This has included investigating C2-MM decision making by varying and controlling demands for levels of conflict resolution, and by observing collaborative mitigation/adaptation responses, considered in terms of coordination (de-conflicting acts), cooperation (de-conflicting means) and collaboration (de-conflicting goals).

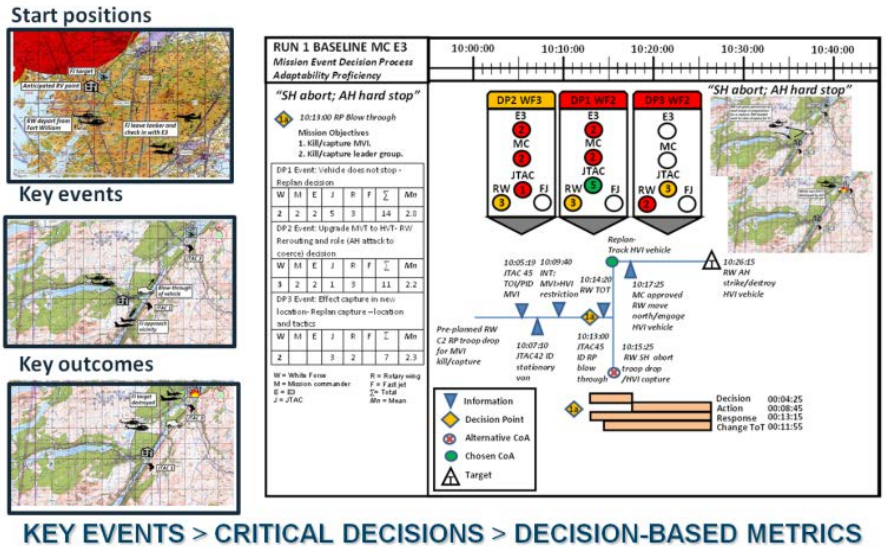

Figure 3: Dstl MCDM Methodology

When designing experiments to measure the effectiveness of HAT in a military context it is important to have militarily representative scenarios which require military operators or autonomous systems to make realistic decisions. Key events must be designed into the scenario to ensure operators are making critical decisions, with effects, in a layered C2 environment, to support our decision based assessment approach.

3.0 HAT METRICS

3.1 Introduction

The technical challenge of proving HAT efficacy, with verification and validation of human-system performance and effectiveness can be broken down in to a number constituent components, the first being the task work performed by a human operator, and the interactions that a human has with other human members of the team. To further add to the HAT system, it is necessary to consider the task work that is undertaken by the autonomy, and the teaming that occurs between autonomous components of the wider team. The final component is the human autonomy teamwork. The combination of the components is illustrated in Figure 4.

Throughout the comprehensive UK MoD Dstl programme of R&D on Autonomy and Mission Systems, a

number of different metrics have been developed predominantly focused on exploiting the critical decision methodology to address the different constituent parts of the complete HAT System. These components are described in the following sections, with the human autonomy teamwork metrics being described in detail.

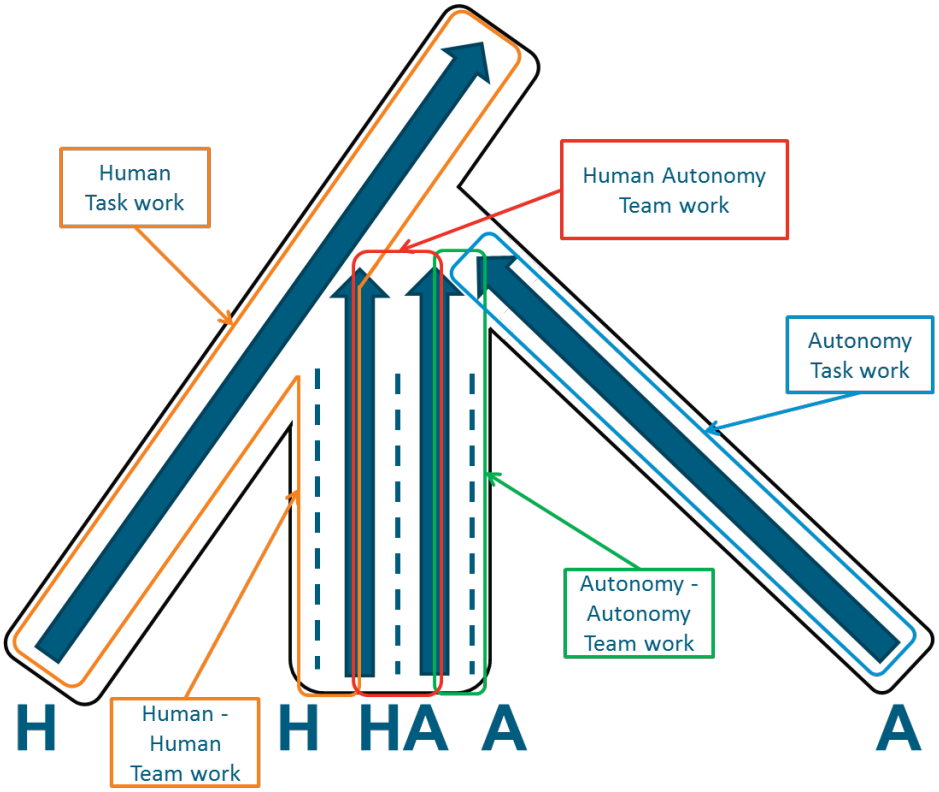

Figure 4: Dstl HAT Component Breakdown

3.2 Human System Teaming

Collaborative Adaptability Proficiency Test Evaluation Assessment Methodology (CAPTEAM) is an independent assessment provided by Dstl, which is designed to estimate mission efficiency by the metrication of Reward and Effort associated with critical mission events and decision processes [3].

The metrics factors measured in CAPTEAM are Workload, Re-plan Task Load, Situational Awareness, Decision Quality, Teamwork, Performance, Tools and Technical System. The metrics are split into Human Task work and Human-Human Teamwork, Task work is assessed in the Participant Protocol and Teamwork in the Subject Matter Expert Observer Protocol using Likert 7-point, Low-High, anchored rating scales.

LOW ANCHORS	DIMENSION Low 1	2	3	4	5	6	High 7	HIGH ANCHORS
WORKLOAD								
Slow/Leisurely/No time pressure	**WL Time Pressure**							*Rushed/Rapid/Frantic*
Very little/Extremely easy	**WL Mental Effort**							*Extremely hard/Stretched*
Relaxed/Placid/Tranquil/Calm	**WL Stress**							*Anxious/Worried/Up tight/Harassed*
REPLAN TASK LOAD								
Forgiving/Simple/Easy	**Replan Decision** Recognise–Evaluate-Mitigate							*Difficult/Demanding/Complex/Exacting*
Forgiving/Simple/Easy	**Replan Action** Disseminate–Acknowledge–Execute-Report							*Difficult/Demanding/Complex/Exacting*
SITUATION AWARENESS								
Stable/Simple/Fixed	**SA Demand on Attentional Resources**							*Unstable/Complex/Variable*
Relaxed/Inattentive/Spare Capacity/Singular	**SA Supply of Attentional Resources**							*Concentrated/Full capacity/Divided*
Uninformative/Meaningless/Unfamiliar	**SA Understanding**							*Informative/Meaningful/Familiar*
DECISION QUALITY								
Indecisive/Doubtful/Guess	**DQ Confidence**							*Decisive /Confident /Obvious*
Dangerous/Vulnerable/Risky	**DQ Survivability**							*Safe/Unthreatening/Secure*
Ineffective/Un-useful/Unproductive	**DQ Effectiveness**							*Effective/Capable/Useful*
Behind/Late/Pressurised	**DQ Timeliness**							*Ahead/On-time/In control*
PERFORMANCE								
Unsatisfactory/Unacceptable/Failure	**Task Performance**							*Perfect/Successful/Satisfied*
Useless/Complicated/Difficult	**Tools Utility**							*Easy/Intuitive/Effective*
No backup plans/ Insensitive/Unresponsive	**Adaptability Proficiency**							*Sensitive/Responsive/Timely effective backup plans*
Fails to meet any mission objectives/0%	**Probability of Mission Success**							*100%/Successfully achieved all mission objectives*

LOW ANCHORS	DIMENSION Low 1	2	3	4	5	6	High 7	HIGH ANCHORS
DECISION QUALITY								
Dangerous/Vulnerable/Risky	**DQ Survivability**							*Safe/Unthreatening/Secure*
Ineffective/Useless/Unproductive	**DQ Effectiveness**							*Effective/Capable/Useful*
Behind/Late/Pressurised	**DQ Timeliness**							*Ahead/On-time/In control*
TEAMWORK								
Silent/Unclear/Confused/Slow/Uninformative	**Communication**							*Clear/Timely/Coherent/Concise/Rapid/ Informative*
Uninformative/Meaningless/Unfamiliar	**Shared SA**							*Informative/Meaningful/Familiar*
Indecisive/Unassertive/Confused/ Uncommunicative	**Leadership**							*Commanding/Directing/Guiding/Initiating management*
Ignoring/Hindering/Blocking/Criticising/ Conflicting	**Support**							*Monitoring/Advising/Correcting/Assisting*
Light/Quiet/Unpressured	**Team Workload**							*Heavy/Busy/Pressured*
PERFORMANCE								
Unsatisfactory/Unacceptable/Failure	**Task Performance**							*Perfect/Successful/Satisfied*
Separate/Independent/Divided/Conflicting	**Collaboration**							*Joint/Shared/Coordinated/Cooperating*
Powerless/Weak/Irrelevant/Not Involved	**Influence Power**							*Influential/Powerful/Forceful/Decisive/Critical*
No backup plans/ Insensitive/Unresponsive	**Adaptability Proficiency**							*Sensitive/Responsive/Timely effective backup plans*
Fails to achieve any mission objective/0%	**Probability of Mission Success**							*100%/ All mission objectives successfully achieved*

Figure 5: Dstl 2011 CAPTEAM Participant Taskwork and Observer Teamwork Assessment Protocols

This assessment methodology relies on military operators to make critical decisions in response to dynamic mission events, with measurable effects, within a realistic military scenario in either a live or synthetic environment. Note that this early 2011 Dstl CAPTEAM protocol was designed for assessing human performance in UK MoD research on Dynamic Air Mission Management (DAMM). Here, the maturity of the automation and autonomy of the technical system was not yet ready to be demonstrated or evaluated as an explicit, discrete capability. In Dstl DAMM research, the focus was on evaluating human effectiveness with distributed networked, C2, at operational and tactical mission levels, with exploitation of human teaming collaboration capability supported by networked mission enabling technologies. The DAMM technical system was treated as being designed to provide tools for human use, in the sense of user-centred mission enabling technologies and operator decision support, with assessment focused on measures of tool usability and utility. Tools usability and utility assessment techniques contrast with the challenges of assessing the efficacy for human use of automatic or intelligent agent technical systems. Such systems are potentially capable of non-deterministic, fully autonomous functioning of operations (sensing, decision making, behaviour), completely independent of human control, with properties such as agency, self-

determination and autonomy associated with advanced Artificial Intelligence (AI), Big Data processing and Machine Learning (ML) technologies.

The Dstl CAPTEAM assessment protocol was used for operator assessments in the 3rd US-UK STRIKE WARRIOR III Joint SE Trial, July 2011 (Cottrell 2011 [4]). Statistical analysis by Dstl of data obtained from application of CAPTEAM indicated the extent to which CAPTEAM component metrics associated with Effort, or Reward are statistically correlated with Probability of Mission Success (PMS) (Taylor and Grabham 2012 [3]). The results showed that relatively beneficial Reward factors, such as improved SA and Decision Quality, were significantly highly correlated with PMS. In comparison, Effort factors and sub-components, such as Workload Stress and Mental Effort a exhibited relatively low, non-significant correlations with PMS. The inter-correlations of the Reward and Effort factors for both Task Work and Team Work in relation to PMS are shown in Figure 6. This illustration uses a structural equation model technique (Castor 2009 [5]). This technique seeks to provide a meaningful, evidence-based, visual data representation of the relationships between the assessment components. The structure of the CAPTEAM components is based on analysis of the data principal components and inter-correlations with PMS The positioning of the sub-components is arranged to be indicative of the progressive strength of the relationships with the intended high level effects. In this graphical representation, improved Adaptability Proficiency and raised PMS are shown to be the outputs, products, or positive effects, arising from the pluri-potential benefits of successful Collaboration in Task Work with Team Work.

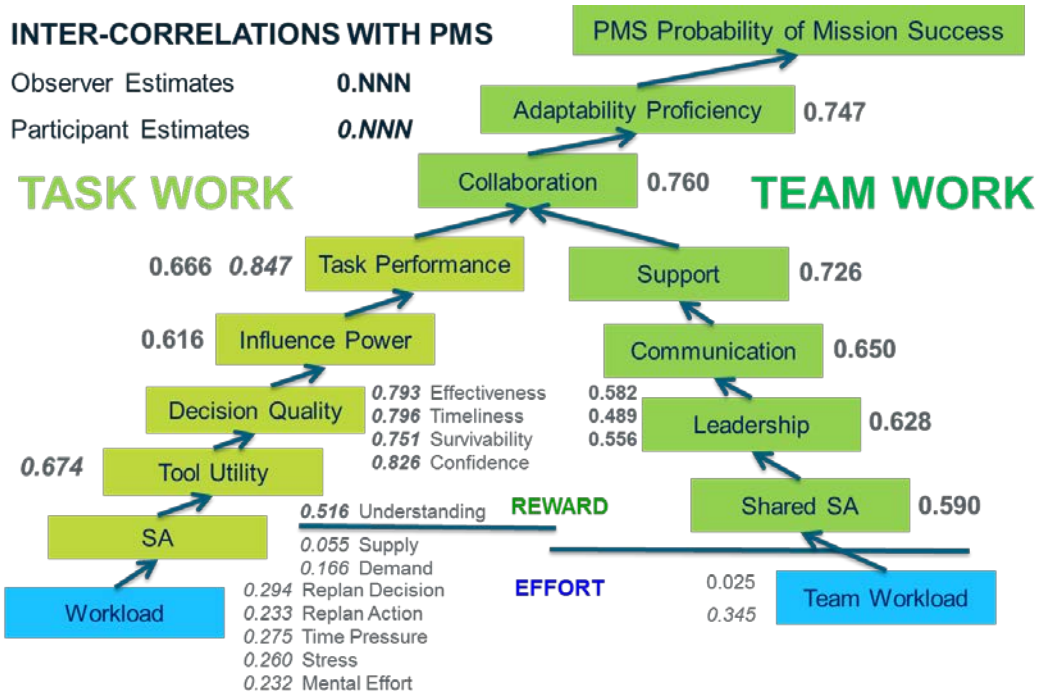

Figure 6: Inter-correlation of the Task work and Team work, Reward and Effort CAPTEAM metrics

An additional benefit from re-using MCDM/CAPTEAM methodology for Autonomous Systems Research T&E consistently over time has been to enable a set of recognised benchmarks to be established, and identified for comparative re-use, in particular for AP and PMS. These benchmarks help frame the assessment context, They act as validated human-system performance standards for comparison, gauging the relative strength of collaborative adaptability proficiency in development prototypes, and providing estimates of human, technical and autonomous system readiness (Taylor 2015 [1]).

3.3 Autonomy Task work

3.3.1 Trustworthiness

Selection of assessment components of autonomy task work seems particularly problematic since much of the detailed functioning of algorithms lacks observability at most commander/operator interfaces. This is reflected in common need for explanation and increased transparency at autonomous systems user interfaces. However, recent work under UK MoD Autonomous Systems Underpinning Research programme, SATLAOP[1] (Searle 2007 [6]) has investigated improving understanding of the requirements for trustworthiness of unmanned systems, based on Publicly Available Specification 754 Software Certification [7]. The results of this work helps guide selection of components of trustworthiness to include safety, reliability, availability, resilience and security components. With the addition of a component for dependability based on human factors literature on trust (Yagoda 2011 [8]), and using the notion of software operation predictability, defined in the sense of adherence to planning in the execution of operations, Dstl have developed and tested both a single and a multi-dimensional subjective rating scale protocol for estimating autonomy task work trustworthiness. The more mature multi-dimensional DSTL HAT Trustworthiness rating scale, with dimension definitions and anchors, is shown below in Figure 7.

HAT TRUSTWORTHINESS ASSESSMENT PROTOCOL								
Low Anchors	DIMENSION							High Anchors
	Low 1	2	3	4	5	6	High 7	
Unreliable; Uncertain; Inconsistent	**Reliability** The ability of the system to operate on missions, perform tasks and deliver effects as specified							Reliable; Certain; Consistent
Undependable; Fickle; Defective	**Dependability** The ability of the system to be relied upon to operate on missions, perform tasks and deliver effects as specified							Dependable; steady; loyal; faithful; responsible; unfailing
Unpredictable; Unlikely; Unexpected	**Predictability** The ability of the system to respond to events and to operate, perform and deliver effects consistently and reliably as planned and anticipated							Predictable; Likely; Expected
Unavailable; Unattainable; Unready	**Availability** The ability of the system to operate on missions, perform tasks, and deliver effects when requested							Available; Attainable; Unready
Brittle; Weak; Slow	**Resilience** The ability of the system to transform, renew and recover in timely response to events							Resilient; Robust; Agile
Unsafe; Unprotected; Unstable	**Safety** The ability of the system to operate without harmful states							Safe; Protected; Stable
Unsecure; Vulnerable; Risky	**Security** The ability of the system to remain protected against accidental or deliberate attacks							Secure; Protected; Assured

Figure 7: Dstl HAT Trustworthiness Protocol

The trustworthiness protocol has been used by the Autonomy Programme in various synthetic environment trials with front line military operators. Data from those trials has shown that operators and software developers are able to complete the protocol understanding the dimensions without difficulty (Taylor, Keirl, Thorpe and Grabham, 2018 [2]). The results indicate that the seven trustworthiness components potentially have useful sensitivity, discriminatory and diagnostic power in assessing HAT. Figure 8 shows the Trustworthiness Ratings obtained on the DAY17 C2 Autonomy Trail from individual software developer SMEs assessing the five individual software component capabilities contributing to the operator UxV GCS system, with comparison WITS Trial Min-Max benchmarks. Data from those trials also indicate that the trustworthiness assessments are not solely dependent on training on a system but on the usability and TTPs associated with the system use.

[1] Self-Aware Trustworthiness Levels and Assurance with Operational Policy

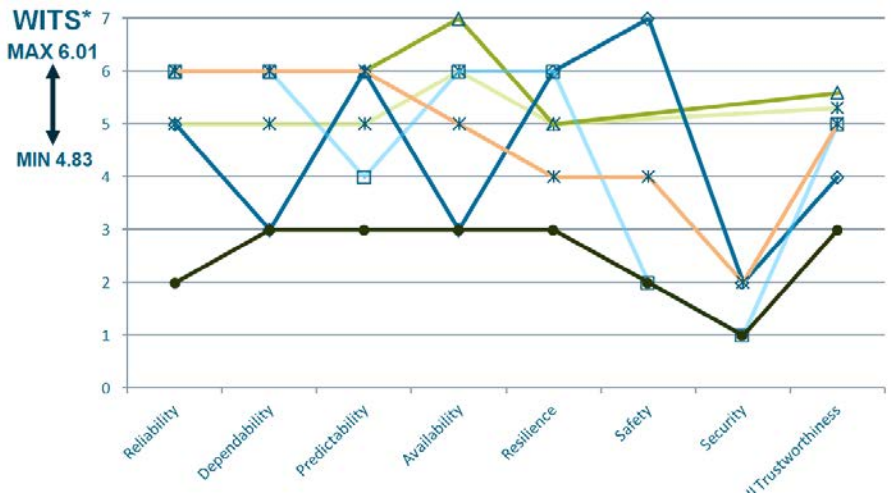

Figure 8: Results of Dstl Trustworthiness Ratings for 5 GCS Software Tools Components from DAY17 C2 Autonomy SE Trial with WITS Trial Comparison Benchmarks

3.4 Human Autonomy Teaming System (Whole System)

3.4.1 Capability Maturity Model

To help guide UK MoD trials in the assessments of autonomy teaming dimensions (human-autonomy and autonomy-autonomy), Dstl have proposed, developed and tested a model with metrics for assessing HAT capability maturity (Taylor, Keirl, Thorpe and Grabham, 2018 [2]). The Dstl HAT Capability Maturity Model (HAT CMM) approach is based on the 1980s software CMM work, originating from Carnegie Mellon University(Carnegie Mellon, 2002 [9]), with consideration of properties of later CMM derivatives, including more people-based dimensions (Systems Engineering, Usability, Organisation, People, Team Process). The underpinning CMMs for Dstl HAT CMM are derived from the basic five software CMM levels (initial through to optimising) with two additional extending, relevant levels, conveniently providing compatibility for integration with the seven level CAPTEAM metrics. The levels progress from Level 0-Not Performed to Level 6-Adaptive, passing through Level 1-Initial, Level 2-Recognised, Level 3-Defined, Level 4-Managed, and Level 5-Optimising. The seven levels of Dstl HAT CMM are each discretely calibrated and defined, with differentiation in terms of goals, processes, and behaviours.

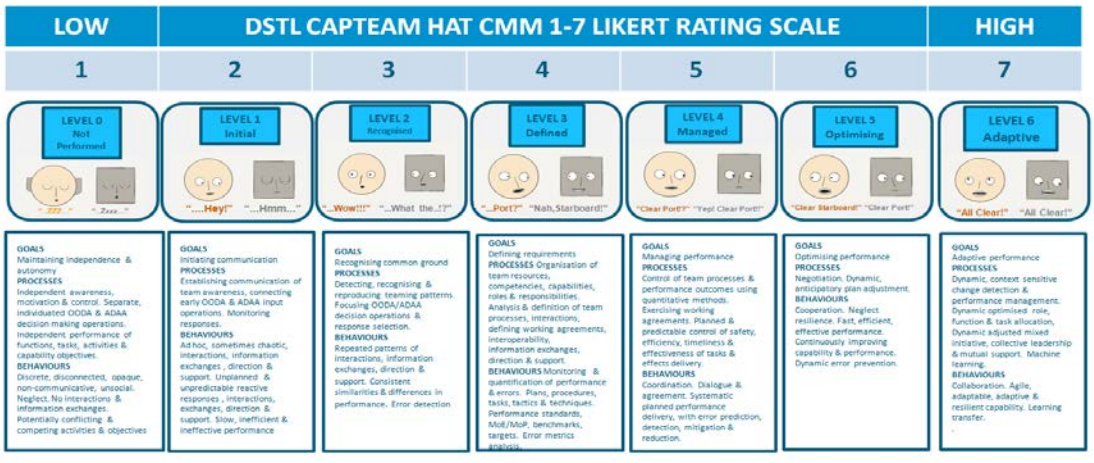

Figure 9: Dstl CAPTEAM HAT Capability Maturity Model Rating Scale

The Dstl HAT CMM Protocol has been applied successfully in recent Dstl Autonomy Research Programme SE and LVC trials. The protocol was used to assess the maturity of UxV GCS technology concepts and systems for assured C2 of Autonomy with military operators. It was found that the maturity assessment concept, including the dimensionality, calibration and scoring method, are novel, complex and needing familiarity and training to understand. Experience and training in relevant and representative military operations and in the proposed technical system use are essential to achieved stability and reliability in HAT maturity identification and classification. Notwithstanding, the capability maturity approach has been found to have both usability and utility for HAT C2 assessment purposes with military operator participant testing in both the DAY 17 SE and CB17Air LVC C2 of Autonomy Trials, The data obtained has indicated that Dstl HAT CMM has the potential to provide both sensitivity and discrimination power, with diagnostic utility, in relation to multiple dimensions of systems readiness and maturity, The scope of potential applications ranges across a diversity of relevant trial and system quality subcomponents, including the maturity of the military operations tested and the maturity of the system mission enabling autonomy technologies.

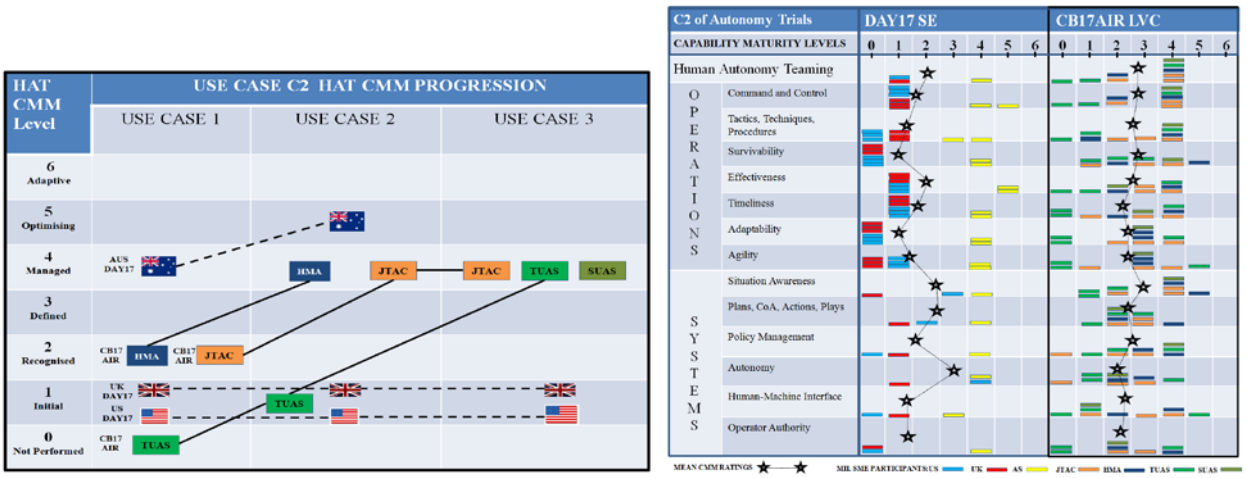

Figure 10: Results of Dstl HAT CMM from DAY17 SE and CB17Air UAS LVC C2 of Autonomy Trials.

3.4.2 Risk Assessment

In Synthetic Environment (SE) and Live, Virtual Constructive (LVC) Trials of Mission Enabling Technologies, assessment methods and protocols typically focus on operator and system performance and effectiveness, necessarily set within the conditions, assumptions, constraints and affordances of the SE/LVC operational trials design environment. To understand the Human Autonomy Teaming system from a broader perspective and identify wider potential requirements and issues, including Defence Procurement Lines of Development (DLODs) that might otherwise be missed, a Dstl Risk Assessment protocol was developed for the WITS[2] trial in March 2017 [10]. The Dstl WITS Risk Assessment protocol was an adaptation of the risk matrix approach used widely in hazard and safety risk assessment. The risk matrix provides identification of calibrated levels of risk probability and impact severity. The Dstl WITS Risk Assessment protocol modification sought to use this framework to address teaming risks for missions. The specific aim of this Dstl Mission Teaming Risk Assessment protocol was to identify the risks of teamwork and the implemented enabling technologies with the focus on the effects on mission success. It is necessary to note, that the capturing of mission teaming risks is focused away from specific platform risks which are typically used in risk based safety assessments of aircraft systems.

[2] Wildcat ISTAR Teaming for Strike, trial in March 2017 was designed to assess the benefits of teaming with potential future platform upgrades

Teaming Risk Assessment Protocol

TEAMING RISK					
NAME..... ROLE/POSITION.....					
DATE..... RUN NUMBER.....					
Mission Critical Decision Teaming Event:					
Description of Risk:					
Mitigation:					
PROBABILITY/ SEVERITY	Extremely Improbable	Improbable	Remote	Occasional	Frequent
Catastrophic					
Severe					
Major					
Minor					
Negligible					

Teaming Risk – Probability and Severity

RISK PROBABILITY OF OCCURRENCE					
Classification	Frequent	Occasional	Remote	Improbable	Extremely Improbable
Quantitative Definition	Likely to occur many times during mission	Likely to occur sometime during mission	Unlikely but possible to occur during mission	Very unlikely to occur during mission	Almost inconceivable to occur during mission

RISK CONSEQUENCE SEVERITY					
Classification	Catastrophic	Severe	Major	Minor	Negligible
Definition	Total failure to achieve mission objectives	Severe reduction in achievement of mission objectives	Major reduction in achievement of mission objectives	Minor reduction in achievement of mission objectives	No effect on mission

Teaming Risk Acceptability Matrix

RISK ACCEPTABILITY					
PROBABILITY/ SEVERITY	Extremely Improbable	Improbable	Remote	Occasional	Frequent
Catastrophic	Review	Unacceptable	Unacceptable	Unacceptable	Unacceptable
Severe	Review	Review	Unacceptable	Unacceptable	Unacceptable
Major	Acceptable	Review	Review	Review	Unacceptable
Minor	Acceptable	Acceptable	Acceptable	Acceptable	Review
Negligible	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable

Figure 11: Dstl Mission Teaming Risk Assessment – Protocol, Definitions and Acceptability Matrix

Risk is defined as a combination of the probability of occurrence and the severity of that occurrence if it should happen. A qualitative probability of occurrence between frequent and extremely improbably and a severity between Negligible and Catastrophic are assigned to each identified risk. The combination of probability and severity results in a matrix of combinations, with catastrophic/frequent the worst outcome and negligible/extremely improbable the best outcome.

The military SMEs (both participants and observers) are required to identify possible risks either that occurred during the scenario or that they could identify as possibly happening as a result of the implemented concepts during the trial. The implemented concepts could either be team orientated, or system/ technology orientated. As a part of the Protocol, the SMEs are able to suggest potential mitigations that may reduce the risk.

Experience with the application of the Dstl Mission Teaming Risk Assessment protocol initially on the WITS trial, and in its development and application subsequently for other Dstl Autonomy SE Research trials, indicate that the protocol purpose is familiar and popular with military operators, it provides a relevant, focusing, useable and useful facility, and a most welcome addition to the Dstl HAT assessment toolkit.

4.0 HUMAN AUTONOMY TEAM WORK - REMEDE

4.1 Introduction

Early Human information processing and team work models were developed in the 1940's, which were expanded upon in the proceeding decades. In the late 1980's MM Taylor developed communication layered protocols which introduced the basis protocol loop. Control abstraction layers were developed in the late 1980's and introduced to support systems for effective C2. All of these approaches model different elements

of human autonomy team work. It is necessary to develop a hybrid model that takes elements from each of the different models to create an assessment methodology that accurately covers Human Autonomy Team work. Each of the different methods will be described in detail, before detailing the hybrid approach that has been developed.

4.1.1 Human Information Processing Phases

In literature on cognitive systems engineering, the functioning of automation is commonly decomposed into a series of four sequential phases of machine information processing, namely Information Aquisition>Information Analysis>Decision Selection >Action Implementation, or AADA (Parasuraman et al, 2000 [11]). This widely used AADA framework mirrors the structure of the simple and popular human decision making OODA loop model (Observe>Orient>Decide>Act) based on understanding of pilot decision making in air combat (Boyd, 1986 [12]; Boyd 2001[13]). The OODA framework, and by implication AADA, has been criticised as a critically deficient over-simplification for conceptualisation of C2, when judged in terms of understanding of control systems principles used in cybernetics (Brehmer, 2005 [14]). This is because the OODA/AADA frameworks fail to identify the functions necessary for effective C2 dynamic decision making, A more complete and efficacious dynamic decision loop, or DOODA loop, includes functions identifying what needs to be accomplished (e.g. information collection, sense-making, and planning), and in particular the provision of feedback on products. Feedback on products is essential for dynamic control and effective C2. The key role in C2 of feedback on performance of critical functions - the products of effective functioning - necessitates representation of the Effects of actions, or Results, in models of C2 dynamic decision making e.g. Information> Decision>Action>Results/Effects.

In CAPTEAM, the Task Work model (Workload>SA>Decision Quality>Task Performance) substantially reflects the OODA Loop structure, with raised AP and PMS as high level Results/Effects, augmented by Teamwork Collaboration. PMS can be further usefully decomposed into attributes of Offensive and Defensive Performance (Castor 2009 [5]). In MOD Air Systems applied research on Mission Enabling Technologies, improved Survivability, Effectiveness and Timeliness (SET) has been long regarded as key performance trade attributes and evaluation assessment criteria for Decision Support Systems (DSS) for combat decision making, It is noteworthy that SET closely mirrors the Cost/Quality/Delivery tradespace commonly used elsewhere in the design of business services. In CAPTEAM, SET are identified as key performance assessment attributes of the Decision Making Task Work function.

4.1.2 Communication Dialogue Protocols

The CAPTEAM Team Work model, based on core attributes of Crew Resource Management (CRM) training, identifies Communication as a key factor in effective Team Work, enabling Shared SA and Leadership Support, and critically underpinning good Coordination, Cooperation and Collaboration. Cybernetics research indicates that communication in natural language, and for effective C2, involves multi-modal dialogue loops between actors with layered protocols for primary messaging and feedback (Taylor, M.M [15]). This dialogue protocol model, with feedback on functioning as a core attribute, applies equally to the design and conduct of effective human-computer or human-autonomy interaction in effective C2 operations.

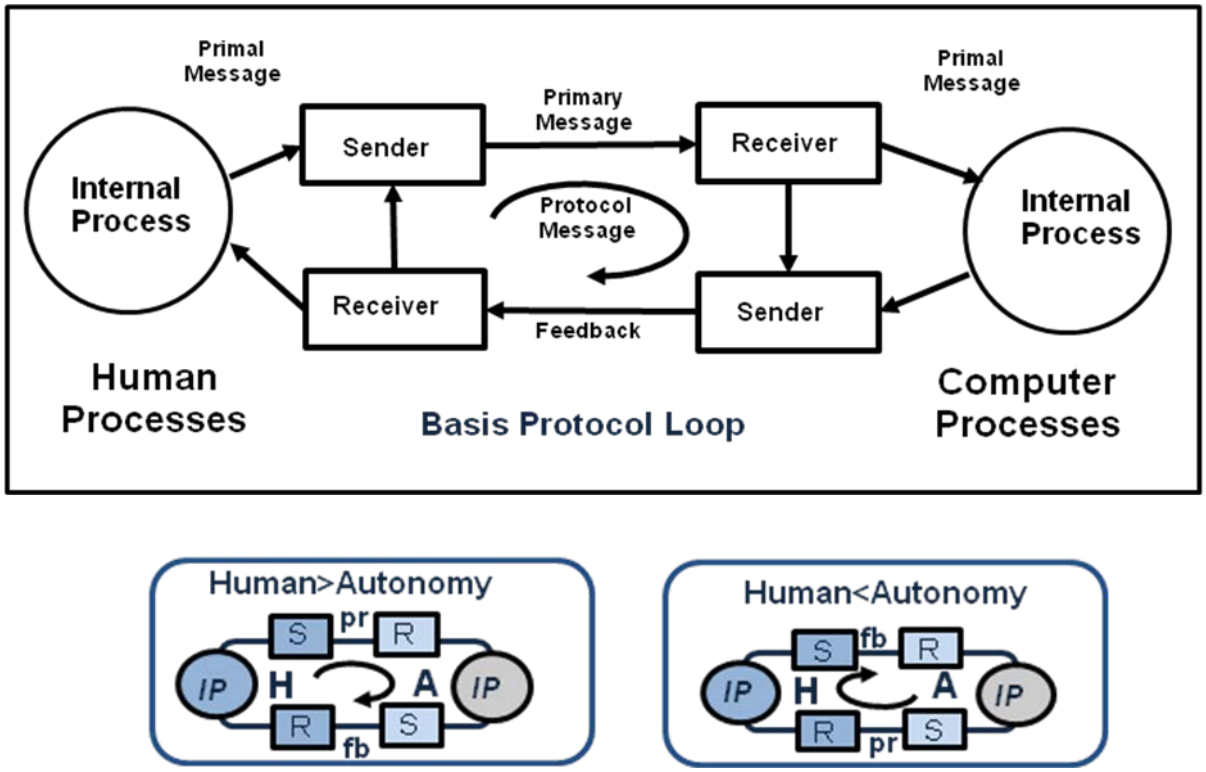

Figure 12: Communication Dialogue Protocols

4.1.3 Control Abstraction Layers

Cognitive systems engineering has shown that systems for effective C2 involve control abstraction layers, modelled largely on the Skills, Rules, Knowledge (SRK), goals-means-acts, hierarchical framework for human behaviour control (Rassmussen [16][17])

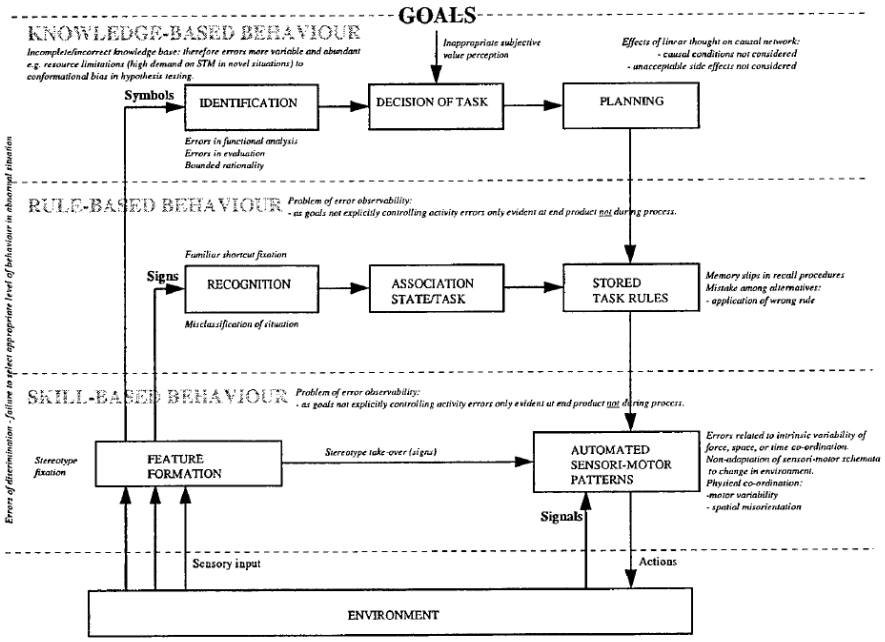

Figure 13: Hierarchical framework for Human behavior control

Subsequent research on Cognitive Work Analysis led to the proposal of a Decision Ladder model with layered components spanning Situation Analysis, Value Judgement, and Planning and Execution: (Observation>Identification>Evaluation > Decision >Planning >Execution) (Vincente, 1999 [18]). The Decision Ladder approach has subsequently been used to represent the collaborative interactions between pairs of actors/agents (Sanderson [19]). This has included control task analysis of user interactions involving PACT levels of automation/autonomy, performed under the DERA Cognitive Cockpit research programme (Taylor, [20], [21], and [22])

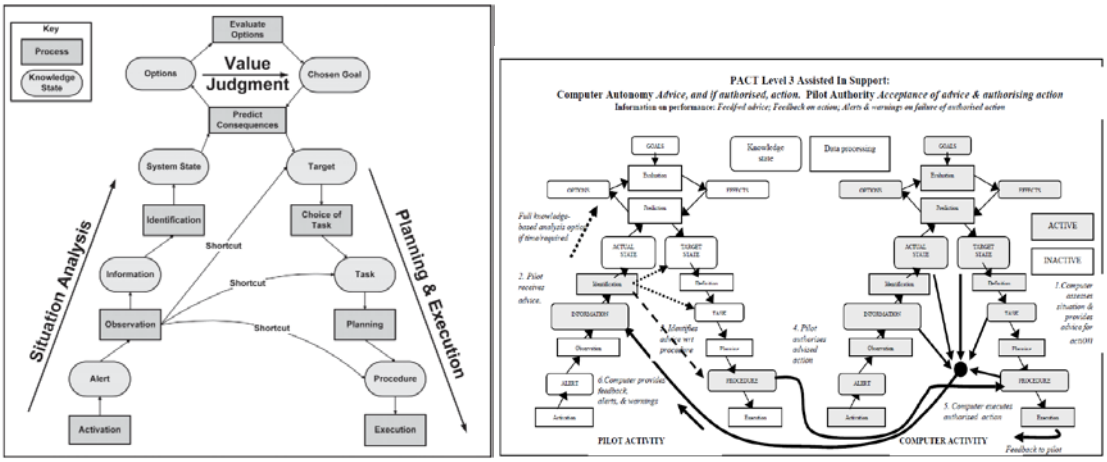

Figure 14: Decision Ladder model and Control Task Analysis of User Interactions

Research on Joint Cognitive Systems (Hollnagel and Woods, 1983 [23], Hollnagel, 1993 [24]) is particularly pertinent to Human-Autonomy Teaming (Taylor 2002 [25]). A multi-layered cognitive control model has been proposed for representation of the control of multiple UxVs (Hollnagel, 2007 [26]). Synchronised,

interacting layers of control identified include Targeting (Governing) > Monitoring (Directing) > Regulating> Controlling (Operating/Tracking).

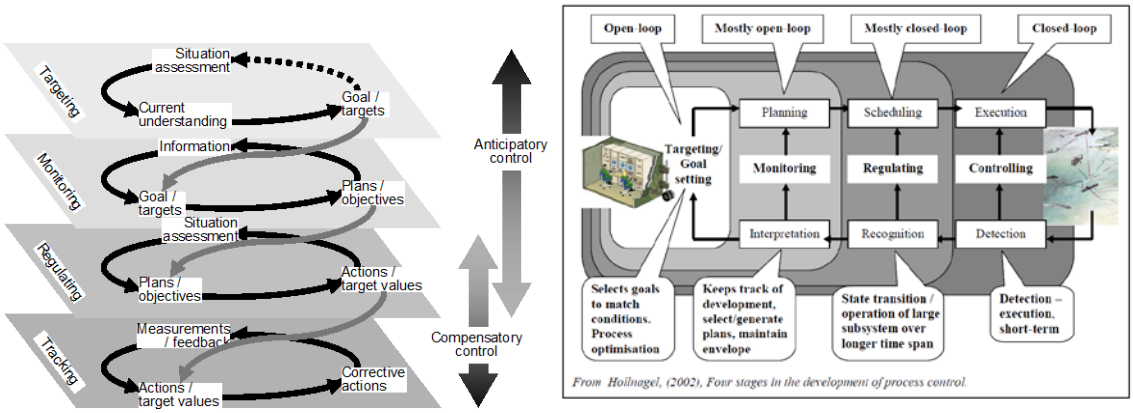

Figure 15: Multi-layered cognitive control model

Informed by this work on cognitive systems engineering and on the concepts of joint cognitive systems, cognitive work analysis and decision ladder control task analysis, work under the Dstl DAMM project identified the applicability of the REMDAER model (Recognise > Evaluate > Mitigate > Disseminate > Acknowledge > Decide > Execute > Report) for representing C2 decision making in a distributed, highly networked military operating environment (Taylor and Grabham 2012 [3]). This framework provided the components for multi-player, distributed or team, decision making cycle, within the operational and tactical C2 architecture C2 OODA or "COODA" layered control system and is illustrated in Figure 16.

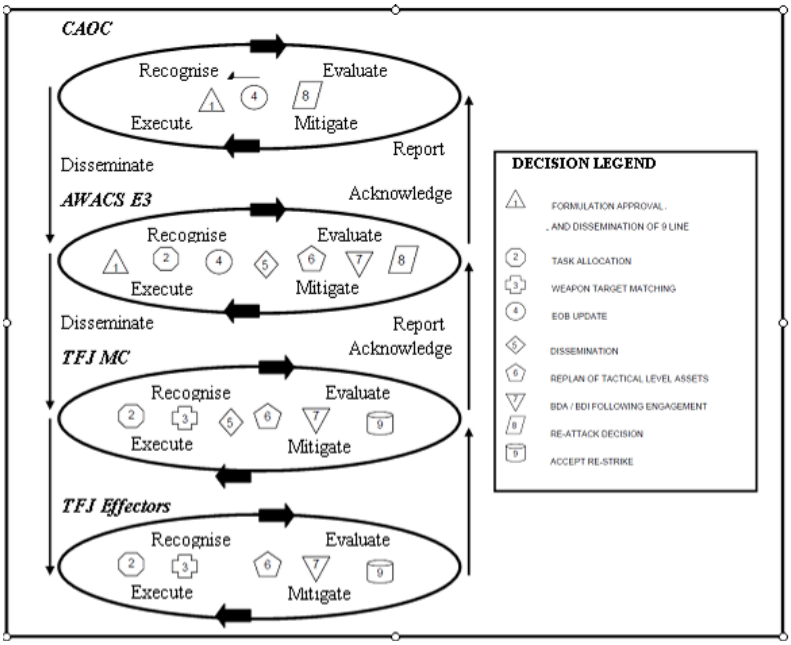

Figure 16: REMDAER Decision Model

4.2 Hybrid Approach

The REMDAER model for the layered control system for airborne mission management was used as an initial step to develop an assessment Protocol for HAT effectiveness. By acknowledging that HAT effectiveness is critically dependent on team member inter-communication, in a similar way to the COODA layered control system, the REMDAER elements were broken down into those that the human and autonomy can complete and those that are dependent on communication between the two. Both humans and autonomy can Recognise, Evaluate, Mitigate and Execute to cause Effects, with Disseminate, Acknowledge and Reporting dependent on the communication Dialogue between the team members.

For assessment purposes, significantly contributing components need to be characteristic and observable; measurable, sensitive and discriminating; and logical, meaningful and evidential. The focus has been on interaction essential information exchange, (and) where the key interactions happen during decision processes, and on the effectiveness of the action taken by the human or the autonomy. In particular the effectiveness of Recognising, Evaluating, Mitigating and Executing (REME) as a team within a mission, and effectiveness of the Communication Dialogue between the human and the autonomy (disseminating, acknowledging and reporting), and the consequences of HAT in terms of specific intended dynamic Effects. Specific measures have been identified to evaluate the effectiveness of the teaming framing around REMEDE.

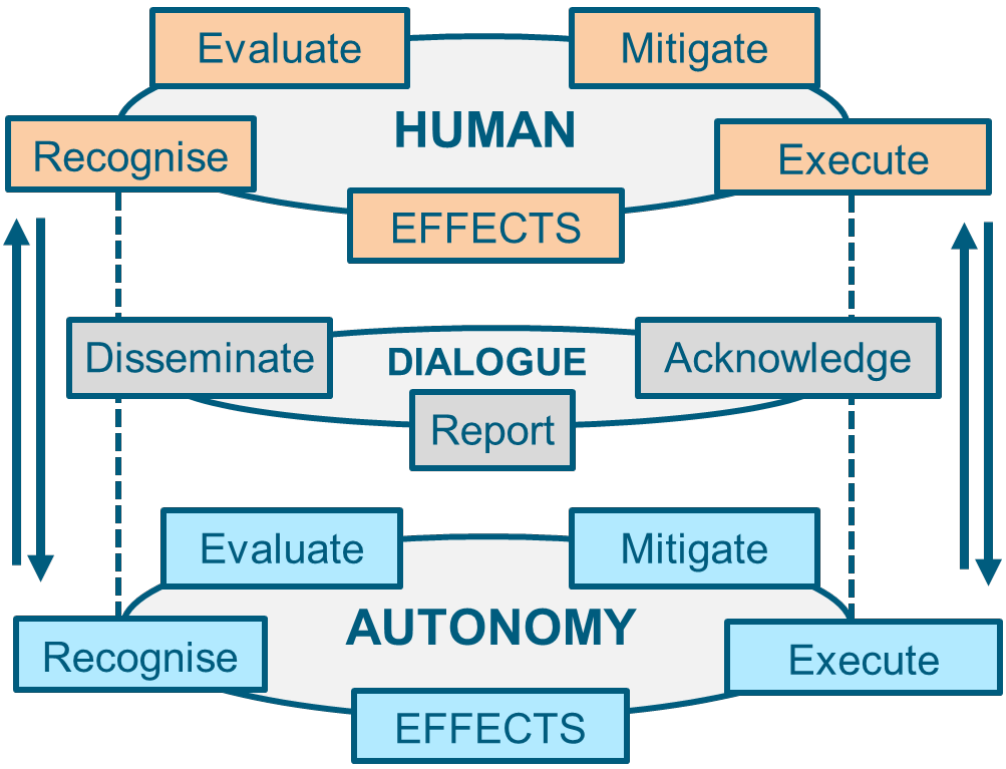

Figure 17: REMEDE Co-Dependent C2 COODA-Model

4.3 REMEDE Assessment Protocols

Seven individual assessment protocols have been proposed initially for assessing the effectiveness of HAT, for validation testing. Six of these protocols are focused on the individual components of the REMEDE model, Recognition, Evaluation, Mitigation, Execution, Dialogue and Effects, with the final protocol being a composite of the REMEDE components. This structure is similar to the Trustworthiness protocol, where the

development and availability of individual protocols and a composite protocol allows for the tailoring of the assessment to the type of experimentation that is being undertaken. Currently neither the composite nor the individual protocols has undergone validation and verification testing, but is scheduled to be utilised in the autumn 2018.

With the exception of the Dialogue protocol, which is unique amongst the different protocols, the remaining six protocols utilise a Likert rating, which ensures compatibility amongst HAT metrics including CAPTEAM, to allow SMEs (operators and/or observers) to score the human functioning component, the system autonomy component and the joint shared communication of each of the REMEDE components. Joint shared communication comprises both the human to autonomy and autonomy to human communication. The Dialogue protocol offers additional granularity for each of the composite REMEDE components in relation to human to autonomy communication and autonomy to human communication, with an individual rating available for each.

EFFECTS	Dstl REMEDE Human Autonomy Teaming Protocol - Effects																			
Impact of actual effects on achievement of Mission and Task Objectives.	Human Functioning Contribution							Joint Shared Communication *H>A & A>H*							System Autonomy Contribution						
	Impact							Impact							Impact						
	Low 1	2	3	4	5	6	High 7	Low 1	2	3	4	5	6	High 7	Low 1	2	3	4	5	6	High 7
Survivability																					
Effectiveness																					
Timeliness																					
Agility																					
Adaptability																					
Offensive Performance																					
Defensive Performance																					
Probability of Mission Success																					
EFFECTS COMPOSITE																					
EFFECTS COMMENTS																					

Figure 18: Dstl HAT REMEDE Effects Component Protocol

5.0 REMEDE VALIDATION AND VERIFICATION TESTING

The REMEDE HAT Assessment Protocol has not yet undergone validation and verification testing However, REMEDE has been adapted for, and will be used in the upcoming STRATUS[3] trials in September 2018 assessing human and system contribution and communication as well as an additional feature asking operators to assess the layered ISTAR contribution towards the REMEDE components. An example assessment protocol from the Stratus evaluation is shown below in Figure 19.

[3] The STRATUS project is assessing the benefits of a layered Intelligence, Surveillance, Target Acquisition, and Reconnaissance (ISTAR) approach in an urban environment with novel autonomy and decision making tools and technologies.

EFFECTS	Dstl REMEDE Human Autonomy Teaming Protocol - Effects																												
Impact of actual effects on achievement of Mission and Task Objectives.	Human Functioning Contribution							Joint Shared Communication							System Autonomy Contribution							Layered C2 ISTAR Collaboration						
	Impact							Impact							Impact							Impact						
	Low 1	2	3	4	5	6	High 7	Low 1	2	3	4	5	6	High 7	Low 1	2	3	4	5	6	High 7	Low 1	2	3	4	5	6	High 7
Raised Understanding																												
Reach Area Depth																												
SA for Specific Effects																												
Survivability																												
Effectiveness																												
Timeliness																												
Agility																												
Adaptability																												
Offensive Performance																												
Defensive Performance																												
Probability of Mission Success																												
EFFECTS COMPOSITE																												
Comments																												

Figure 19: Dstl HAT REMEDE Effects Stratus Specific Component Protocol

6.0 CONCLUSIONS

The assessment of HAT efficacy, while technically challenging can be broken down into a number of individual components of the Human Autonomy Teaming System to simplify the challenge. The understanding of the various components, Human Task work and Teamwork, Autonomy Task work and Teamwork and Human Autonomy Teaming, of the HAT system has been built up over more than 18 years of UK MoD Dstl R&D on Autonomy and Mission Systems. While it is possible, and in some cases easier to understand the HAT system in its entirety, it is often not possible to identify the locus of success or failure at such a high level depending on the system that is being tested.

The Human task work and teamwork elements of the HAT system, through CAPTEAM which is designed to estimate mission efficiency by the metrication of Reward and Effort associated with critical mission events and decision processes. CAPTEAM has been extensively utilised for trials within the UK research programme, with statistical analysis indicating the inter-correlation of the assessment components with relatively beneficial Reward factors, such as improved SA and Decision Quality, were significantly highly correlated with PMS. In comparison, Effort factors and sub-components, such as Workload Stress and Mental Effort a exhibited relatively low, non-significant correlations with PMS.

The Autonomy Task Work component of the HAT system is addressed through a broken down multi-dimensional Trustworthiness scale. While only used in a limited number of synthetic trials, the results indicate that the seven trustworthiness components potentially have useful sensitivity, discriminatory and diagnostic power in assessing HAT, with data from the trials also indicating that trustworthiness is not solely dependent in training on a system but on the usability and TTPs that are associated with the system in use.

The entirety of the HAT system can be assessed through a combination of the HAT Capability Maturity Model, which was derived and developed from software CMMs developed by Carnegie Mellon University and the Risk Assessment. The HAT CMM has been used successfully in recent Dstl Research Programme SE and LVC trials, assessing UxV technology concepts and systems for assured C2 of Autonomy with military operators. It was found that the concept of the maturity assessment, are novel, complex and needing familiarity and training to understand. With experience and training in relevant and representative military

operations and in the proposed technical system use are essential to achieve stability and reliability in HAT maturity identification and classification. Notwithstanding, the approach to capability maturity has been found to have both usability and utility for HAT C2 assessment purposes, with the obtained data indicating that the Dstl HAT CMM has the potential to provide both sensitivity and discrimination power. The Risk Assessment protocol has been used in both synthetic and LVC trials recently, and provides a broader perspective of the HAT system and allows for the potential identification of wider requirements and issues, including those associated with the Defence Procurement Lines of Development that might otherwise be missed. The focus of risks is away from the traditional platform safety risks that are typically addressed in risk based safety assessments of aircraft systems, but instead focused on the risks associated with teamwork and the implemented enabling technologies with respect to mission success.

The new component of the UK HAT System assessment methodologies is the REMEDE assessment protocol. The protocol itself is focused on the Human Autonomy Teamwork component of the entire system, and has been developed from a variety of information processing, communication and team work models. Previously these had been combined to form the REMDAER decision model, as part of previously reported work for the DAMM project, which was focused on a multi-player, distributed or team, decisions making cycle within the operational and tactical C2 architecture "COODA" layered control system. With the critical dependency of team member inter-communication being similar across the HAT system and the REMDAER layered control system it is possible to break down the REMDAER elements into what the Autonomy can complete and what the Human can complete and what dialogue is required between them to be effective. This results in the components of Recognise, Evaluate, Mitigate, Execute to cause effects being applicable to the Autonomy and the Human, with Dissemination and Reporting and Acknowledging forming the dialogue between them. The application of this to the assessment of HAT has been proposed in two different ways, from a high level composite protocol that only looks at high level of the REMEDE components, to an individual protocol for each of REMEDE components. This breakdown of the individual components allows for an increased granularity of the underlying aspects of the different components. At this stage, the REMEDE protocol has been selected for use with the STRATUS project under the UK MoD Dstl Autonomy Research programme, with a specific instantiation developed for the specifics of the trial. This trial will serve as the initial opportunity to undertake verification and validation testing of the REMEDE protocol, the outcomes of which will allow a greater understanding of the discrimination and sensitivity of the data captured.

Autonomy Teamwork has not been addressed within the UK research programme at this time and remains an area of interest in the context of Human Autonomy Teaming and the contribution that Autonomy Teaming has to the understanding the efficacy of the entirety of the HAT System.

7.0 REFERENCES

[1] Taylor R.M. 2015. Human Autonomy Teaming: Verification, Benchmarking and Readiness. In, RPAS Achievements and Challenges. Proceedings of the 2015 Presidents Conference, Royal Aeronautical Society, London' 7-8 October 2015. SG UAS Event Code 772. ID P2PP2R-2015-09-28T 10:34:36. DSTL/TR91332 967705572. ISBN 1 8576831X.

[2] Taylor, R.M., Keirl, H., Thorpe E. and Grabham A. 2018. UK-1: HAT-CV&V Human Autonomy Teaming – Collaboration Verification and Validation. Final Report NATO HFM247. November 2017.

[3] Taylor R.M. and Grabham A. 2012. UK-1: Dynamic Airborne Mission Management. In, Supervisory Control of Multiple Uninhabited Systems – Methodologies and Enabling Human-Robot Interface Technologies. Chapter 13, pp 1-27. Lessons Learnt. Annex A, pp 1-14. HFM-TG-170. AC/323(HFM-170)TP/451. December 2012

[4] Cottrell, R.J. 2011. RE314 July 2011 Strike Warrior III Joint Experimental Trial Report. QinetiQ/AEG/TI/CR1102929/V0.1. 14 Dec 2011. Cody Technology Park, QinetiQ Ltd., Farnborough, UK.

[5] Castor, M. 2009. The use of structural equation modelling to describe the effect of operator functional state on air-to-air engagement outcomes. Linkoping Studies in Science and Technology. Dissertation No. 1251. Linkoping University Institute of Technology. Linkoping, Sweden. ISBN 978-91-7393-657-6 ISSN 0345-7524.

[6] Searle T. 2017. ASUR Self-Aware Trustworthiness Levels and Assurance with Operation Policy 1014_C3_Ph2_060/073/110, 49048-45595R Issue 2, Frazer-Nash Consultancy Ltd.

[7] British Standards Institution, PAS 754:2014, 2014.

[8] Yagoda, R.E. 2011. WHAT! You want me to trust a ROBOT? The development of a human robot interaction (HRI) trust scale. MSc. Psychology Graduate Thesis. North Carolina State University, Raleigh, North Carolina 2011.

[9] Carnegie Mellon. 2002. Capability Maturity Model® Integration (CMMISM), Version 1.1. CMMISM for Software Engineering (CMMI-SW, V1.1). Continuous Representation. CMU/SEI-2002-TR-028. ESC-TR-2002-028. CMMI Product Team, August 2002. Carnegie Mellon Software Engineering Institute, Pittsburgh, PA 15213-3890

[10] Keirl, H, Hennessy, DKR, Thorpe, A, Taylor, R, Halpin, D. 2017. Wildcat ISTAR Teaming For Strike – Phase 1 Experimentation. DSTL/CR101870 Issue 1. March 2017. Defence Science and Technology Laboratory, UK MoD.

[11] Parasuraman, R., Sheridan, T. and Wickens, C. 2000. A Model for Types and Levels of Human Interaction with Automation, IEEE Transactions on Systems, Man, and Cybernetics. – Part A: Systems. and Humans, Vol. 30, pp. 286-297.

[12] Boyd, J.R. 1986. Organic Design for Command and Control. Briefing slides, October 1986

[13] Boyd, J.R. 2001. “An essay on winning and losing”. http://defence and the national interest. d-n-i-net, 2001

[14] Brehmer, B. 2005. One Loop to Rule Them All. In Proceedings of 11TH ICCRTS Coalition Command and Control in the Networked Area: C2 Analysis, C2 Modelling and Simulation, Cognitive Domain Issues. Department of War Studies, Swedish National Defence College, Stockholm, Sweden.

[15] Taylor M.M. 1989. Response Timing in Layered Protocols: A Cybernetic View of Natural Language. In, The Structure of Multi-Modal Dialogue. M.M. Taylor, F. Neel, and D.G.Bouwhuis (Eds). Chapter 11, pp 159-172 Amsterdam: Elselvier Sciences B.V. North Holland.

[16] Rasmussen J. 1986. Information Processing and Human- Machine Interaction: An Approach to Cognitive Engineering. New York: North Holland.

[17] Rasmussen, J., Pejtersen, A.M., and Goodstein, L.P. 1994. Cognitive Engineering: Concepts and Applications. New York: Wiley.

[18] Vicente, K.J. 1999. Cognitive Work Analysis: Towards Safe, Productive, and Healthy Computer based

Work. New Jersey: Lawrence Erlbaum.

[19] Sanderson, P., Naikar, N., Lintern, G., and Goss, S. 1999. Use of cognitive work analysis across the system life cycle: From requirements to decommissioning. Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting, Houston, TX. Santa Monica: HFES

[20] Taylor, R.M. 2001. Cognitive Cockpit Control Task Analysis. DERA Memo, DERA/CHS3/6.3/14/7, 07 March 2001.

[21] Taylor R.M 2002. Technologies for Supporting Human Cognitive Control. In Human Factors in the 21st Century. NATO MP-077. 18:1-14. RTO Human Factors and Medicine Panel (HFM) Specialists' Meeting, Paris, France, 11-13 June 2001.

[22] Taylor, R.M 2007. Human Automation Integration with Contractual Autonomy. In, Uninhabited Military Vehicles (UMVs): Human Factors in Augmenting the Force. Final Report of RTO Human Factors and Medicine (HFM) Task Group HFM-078/TG017. Chapter 7, Human Automation Integration, pp 3-12 to 29. AC/323(HFM-078)TP69. ISBN 978-92-837-0060-9 July 2007.

[23] Hollnagel, E. and Woods, D. D. 1983. Cognitive Systems Engineering: New wine in new bottles, International Journal of Man-Machine Studies, 18, pps

[24] Hollnagel E. 1993. Human Reliability Analysis: Context and Control. London: Academic Press. 583-600.

[25] Taylor, R.M. 2002. Capability, Cognition and Autonomy. RTO HFM Symposium on "The Role of Humans in Intelligent and Automated Systems", Warsaw, Poland, 7-9 October 2002, NATO RTO-MP-088. KN3, 1 to 26.

[26] Hollnagel, E. 2007. Modelling Multi Layered Control: Application of the Extended Control Model to the Analysis of UAV Scenarios. In, Uninhabited Military Vehicles (UMVs): Human Factors in Augmenting the Force. Final Report of RTO Human Factors and Medicine (HFM) Task Group HFM-078/TG017. Chapter 7.7, Human Automation Integration, pp 73 to 96. AC/323(HFM-078)TP69. ISBN 978-92-837-0060-9. July 2007.