Modeling Operator Performance through Task-Oriented Machine Learning

Bryan Vandrovec
RED-INC
Human Systems Division
California, MD
bryan.vandrovec@red-inc.us

Robert Lutz
The Johns Hopkins University
Applied Physics Laboratory
Laurel, MD
robert.lutz@jhuapl.edu

Timothy Bagnall
Mosaic ATM
Leesburg, VA
tbagnall@mosaicatm.com

Tracy Sanders
The MITRE Corporation
McLean, VA
tlsanders@mitre.org

ABSTRACT

Autonomous systems are quickly evolving to provide a versatile and essential capability in both military operations and commercial applications. From a human-systems perspective, these recent technological developments are changing the role of human operators into that of supervisory controllers of complex automated and autonomous systems who must maintain situation awareness (SA), and be ready to rapidly intercede in complex or critical situations that require human judgment and general intelligence. Unfortunately, this rapidly advancing technology has exceeded the ability of traditional methods, often relying on expertise and intuition, to predict how operators will perform and interact.

In support of U.S. Navy unmanned aircraft system (UAS) airspace integration initiatives, a high-fidelity simulated representation of air vehicle operator (AVO) behavior and performance is in development. The Operator Model (OM) employs machine learning (ML) and other artificial intelligence techniques for characterizing observed responses of AVOs to air traffic encounters, along with a means to reproduce and generalize those responses for use in faster-than-real-time constructive computer simulations. In doing so, this model addresses several needs, such as providing an economical means of generating the volume and variety of human-performance data required for platform certification, and informing future design and training decisions.

The OM represents a significant new capability for unmanned aviation-systems development. It combines task analysis and experimental psychology with advances in machine learning to support simulation-based acquisition in a complementary and cost-effective manner, enabling certification of defense systems with higher levels of autonomy and more complex patterns of human-computer interaction (HCI). This paper will provide an overview of the OM hardware and software architecture, and highlight the Live-Virtual-Constructive (LVC) trials that have been performed at the Naval Air Warfare Center, Aircraft Division (NAWCAD) to validate UAS sense-and-avoid (SAA) capabilities. Phase 1 results indicate a strong agreement between LVC and OM measures of effectiveness (MOEs).

ABOUT THE AUTHORS

Bryan Vandrovec is a Principal Software Architect at Research and Engineering Development, LLC (RED-INC). Mr. Vandrovec graduated from Cornell University's School of Engineering with a B.S. in Computer Science, and earned an M.S. in Systems Engineering from Johns Hopkins. He is currently supporting the U.S. Navy on MQ-4C Triton Airspace Integration as the AIR-4.6 Technical Lead for the design, development, and deployment of the Human Performance Modeling and Simulation (M&S) capability, which includes the Operator Model and Data Collection and Management System. In this role, he provides technical management and software and systems engineering expertise, including tailoring of process models, requirements analysis, risk management, object-oriented analysis and design (OOAD), and database and machine-learning development. Mr. Vandrovec also serves as Project Integration Coordinator for transition and integration of the Operator Model to AIR-5.4 and FFRDC partners in accordance with the overarching Triton Airspace Integration M&S Plan. Prior to his work on Triton, Mr. Vandrovec served as a subject matter expert (SME) for DARPA's Collaborative Operations in Denied Environments (CODE), and supported the

Navy's Common Control System (CCS) and P-8A Poseidon programs in the area of mission systems architecture. He is also an inventor with several patents, and was a co-founder of Immersive Technologies, LLC.

Timothy Bagnall is a Principal Analyst with Mosaic ATM in Leesburg, VA. Mr. Bagnall earned an undergraduate degree in Systems Engineering with a minor in Economics from the University of Virginia. He has formal training in systems engineering, lean management, operations research, probabilistic systems, risk analysis, intelligent decision systems, and modeling and simulation. His professional career has involved applying his expertise in systems engineering, modeling and simulation, and lean management to provide a wide range of solutions for the DoD and industry. He has served as an expert consultant of human performance modeling and human factors engineering for the Army, Navy, Air Force, and NASA. Highlights from his career include an Air Force maintenance software program that predicts the impacts of human performance on mission capability and readiness; an attention and situational awareness human performance model to provide an analysis of alternatives of displays for wake vortices in aircraft cockpit displays; the design and development of an automated patient flow performance reporting system for a major healthcare system; and the design of a hospital occupancy simulation for a major healthcare system that saved millions of dollars through the right-sizing of a proposed hospital addition.

Tracy Sanders is a senior systems engineer with The MITRE Corporation in McLean, VA. Dr. Sanders earned an M.S. in Modeling and Simulation and a Ph.D. in Applied Experimental and Human Factors Psychology from the University of Central Florida. Her expertise in human factors, trust in automation, unmanned systems, usability, and interaction design has been leveraged on projects with organizations including NASA, DARPA, the FAA, the FCC, the Army Research Laboratories (ARL), and the Veteran's Health Administration (VHA). Currently, Dr. Sanders supports the Navy's MQ-4C (Triton) Program by providing human factors guidance, experimentation, and model validation efforts.

Robert Lutz is a principal staff scientist at The Johns Hopkins University Applied Physics Laboratory in Laurel, MD. His background includes 37 years of practical experience in the development, use, and management of models and simulations across all phases of the DoD systems acquisition process. He currently serves as the Navy's MQ-4C (Triton) Program M&S lead in the Airspace Integration area. He also supports LVC testing for several autonomy science and technology (S&T) programs, such as the Safe Testing of Autonomy in Complex Interactive Environments (TACE) Project and the Collaborative Operation in Denied Environments (CODE) Program. In addition, Mr. Lutz serves as the Chair of the Simulation Interoperability Standards Organization (SISO) Board of Directors and Vice Chair of the SISO Executive Committee; serves on the Tutorial Board and Fellows Committee at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC); and he is a guest lecturer on various M&S-related topics in The Johns Hopkins University Whiting School of Engineering.

Modeling Operator Performance through Task-Oriented Machine Learning

Bryan Vandrovec
RED-INC
Human Systems Division
California, MD
bryan.vandrovec@red-inc.us

Robert Lutz
The Johns Hopkins University
Applied Physics Laboratory
Laurel, MD
robert.lutz@jhuapl.edu

Timothy Bagnall
Mosaic ATM
Leesburg, VA
tbagnall@mosaicatm.com

Tracy Sanders
The MITRE Corporation
McLean, VA
tlsanders@mitre.org

BACKGROUND

All of the major military Services have assigned roles and responsibilities. The U.S. Navy is responsible for maritime supremacy, and serves as the guardian of oceans, waterways, and littoral areas throughout the world. The need to conduct surveillance of the maritime environment in an effective and efficient manner is key to the achievement of supremacy. The P-3 Orion has performed this mission admirably for decades, with the P-8A Poseidon now transitioning into this role. However, the long on-station times required of manned surveillance platforms as well as advancements in unmanned aircraft systems (UAS) technology have resulted in the introduction of the MQ-4C Triton UAS into the Fleet (see Figure 1). Triton is a variant of the U.S. Air Force's RQ-4B Global Hawk UAS optimized for the maritime intelligence, surveillance, and reconnaissance (ISR) mission. Triton can also be deployed in support of missions such as battle damage assessment (BDA), port surveillance, and communication relay (U.S. NAVY, 2017).

Figure 1: U.S. Navy MQ-4C (Triton)

In order for Triton to perform its assigned missions, it must be able to operate within airspaces containing vastly different characteristics. In future missions, Triton must be able to operate in relatively low-density oceanic environments, higher-density offshore regions, and even higher-density overland/littoral areas within the national airspace system (NAS), including the terminal environment. In all these different airspaces, a UAS must achieve an equivalent level of safety as compared to a manned aircraft system. However, many of the policies and procedures used to enforce flight safety standards as mandated by the Federal Aviation Administration (FAA) and the U.S. Department of Defense (DoD) need to be modified and extended to account for unmanned aircraft system (UAS) operations. Similarly, there is a need for the International Civil Aviation Organization (ICAO) to extend policies and rules for civil UAS operations in international airspace. While such work is ongoing, the execution of the Triton deployment schedule requires a formal certification that Triton can operate safely for its Initial Operational Capability (IOC) missions and deployment areas. The basis of this certification is a *safety case*, as discussed below.

As part of the overall certification plan, Triton Airspace Integration has adopted a tailored version of the standard V-model for system verification and validation (V&V). A key element of this process is assessing the performance of the air vehicle operator (AVO) in the context of self-separation and collision avoidance. In support of this objective, a human-systems support effort is underway to conduct live-virtual-constructive (LVC) experiments as a means of complementing fast-time constructive simulations for evidence generation. Informed by the preceding constructive phase, the LVC trials are gathering crucial data aligned to task measures of performance (MOPs) and providing the basis for development of an empirical *Operator Model* to generalize that response within a series of follow-on

constructive runs. In this way, the practical limitations of traditional human-performance studies and flight testing are mitigated, enabling the program to gather the volume of evidence needed to determine risk ratios[1] and ensure Triton will achieve its target level of safety (TLS)[2] in the operational environment.

NAVAIRINST 13034.4 (NAVAIR, 2014) defines a "safety case" as:

> *The process by which a formally documented body of evidence is created that provides a convincing and valid argument that a system is safe for a given application in a given environment. The safety case documents the safety requirements for a system, provides evidence that the requirements have been met, and documents the argument linking the evidence to the requirements. Elements of the safety case include safety claims, evidence, arguments, and inferences.*

The traceable chain of reasoning arises from the decomposition of top-level safety claims to sub-claims, to the level that sub-claims can be supported directly by evidence. This results in a tree structure terminating in evidence items supporting sub-claims. If the evidence is valid and complete, and the claim decomposition sound, then the top-level claim is supported (Lutz, 2017).

NAVAIR is using a safety-case approach to direct the collection, organization, analysis, and presentation of the Triton certification rationale to the certification authority. Due to the high costs and technical challenges of replicating the full complexity of the Triton operational environment on a live range as well as the safety concerns associated with direct Triton interaction with live test assets, a considerable amount of the evidence for certification is being produced via modeling and simulation (M&S). Since M&S results are to be used to directly satisfy evidence requirements, such results must be highly credible. The generation of statistically significant data requires the application of sophisticated design-of-experiments (DOE) techniques along with Monte Carlo analysis methods that employ faster-than-real-time (fast-time) constructive simulation tools, such as IMPRINT Pro. M&S credibility is demonstrated via a structured V&V process and supporting test plans/tools and that together provide the necessary accreditation evidence for both the individual M&S tools and the integrated M&S environment. NAVAIR committed significant resources to the Triton safety case analysis to ensure the efficacy of M&S results from both of these perspectives.

Since human operators are a critical component of any unmanned system, the fast-time tools must accurately represent the behaviors and performance of the Triton AVO in order to accurately predict the level of safety for the entire Triton system. In early Triton safety-case analysis, a set of Triton AVO-validated "if-then" pilot rules that collectively defined what a Triton AVO should do under different operational conditions was developed and integrated into the fast-time tools. However, this representation of the human AVO provided an overly optimistic view of total system safety performance. The need for an innovative approach to address this deficiency led to the Operator Model (OM) approach discussed throughout this paper. That is, the application of machine learning and other artificial intelligence techniques and leveraging actual AVO behaviors and performance observed during immersive LVC exercises to produce a highly realistic AVO representation for use in the fast-time constructive simulation tools.

Aligned to the safety case, and identified through a task analysis (discussed below) informed by use cases from the Broad Area Maritime Surveillance – Demonstrator (BAMS-D), measures of performance (MOPs) associated with the duration and output of each AVO task have been identified and documented in a traceability matrix. These metrics include signal detections, accuracy of threat evaluations, correctness of action decisions and maneuver commands, and their corresponding response times. Appropriate data elements produced by the LVC apparatus were then identified and tied to each MOP, and a set of custom trace diagrams expressed in the Unified Modeling Language (UML) were created to define the MOP evidence chains. To obtain the MOP-aligned data from the LVC trials, an extensive data-collection, management, and analysis infrastructure was developed to support configuration and training of the OM, as detailed in the sections that follow. The Operator Model itself was developed using a custom interdisciplinary process partially derived from ICONIX, with best practices in software/human-factors engineering.

---

[1] The risk ratio, also referred to as relative risk, is a quantitative measure of the relative benefit of equipping with a system when compared to a reference system or situation. Risk ratio is the primary analysis metric used to assess the safety performance of the Traffic Alert and Collision Avoidance System (TCAS).

[2] A generic term representing the level of risk considered acceptable in particular circumstances (ICAO, 2010).

APPLICATION OF MACHINE LEARNING

The theoretical basis employed for the application of machine learning on the OM was primarily derived from the concepts and principles presented in *Learning from Data* (Abu-Mostafa, 2012). Care has been taken to ensure best practices such normalization of heterogeneous state vectors are observed prior to the application of various types of learning algorithms. A precept central to OM development was "let the data decide."

Design Considerations

The following characteristics were among those considered as design drivers in the selection of suitable machine-learning algorithms and approaches for use in OM task aspect models (TAMs) (discussed in the Architecture section):

- Fast training to allow for more solution iterations
- High accuracy on relatively small data sets
- Insensitivity to outliers – "no outliers in human factors"
- Generation of posterior probability scores (classifiers)
- Approaching Bayes optimal classification under uncertainty (noise, small sample size, etc.)
- Universal approximation on compact subspace in $\mathbb{R}^n$
- Preservation of stochastic behavior (w/ cluster pre-process)

Additionally, analysis of task-separated data frames informed the selection of TAM strategies (see Figure 3), particularly in regards to decision tasks, with deterministic classifiers suitable for highly separable data, and partially separable data calling for non-deterministic implementations to randomly select among multiple options based on their context-dependent likelihood. Moreover, heuristic overlays were often applied to support enforcement of domain constraints on the outputs. An example of this is the quantization of altitude override commands to multiples of 100, 500 or 1000 feet above certain magnitude thresholds. Where insufficient training data were obtained, and concise domain rules were available from subject matter experts (SMEs), a fuzzy-logic inference system (FIS) was employed. For the Triton Phase 1 data, this technique was limited to the output TAM of the evaluate threats task (see Task Analysis section) as the LVC trials afforded only seven examples of near-miss encounters. Alternatively, augmenting heuristics were sometimes used to bolster the machine-learning elements, particularly in regions of the feature space that were sparsely populated with empirical data. In other cases, simple probabilistic models based on best-fit probability density functions (PDFs) were deemed sufficient by regression or cluster analysis as the observed operator response did not show a significant dependency on elements from the world-state vector. The calculation of AVO detect/observe times is such a case.

Task Network Embedding

Embedding the TAM machine-learning elements within a task network provided both a mechanism for composing arbitrarily complex behaviors and a familiar structure to support more traditional methods of human-factors analysis. While helping to address a commonly cited drawback of machine-learning systems, namely opacity, this arrangement does present a challenge to the convergence of model behavior with that observed in the field or laboratory. Because of this, in addition to the shared situation awareness (SA) construct, a select set of tuning parameters was defined within the model (e.g., threat-detection sensitivity) to allow for top-level fitting of the model response to the empirical data. Ideally, this tuning would be accomplished automatically by means of another level of machine learning. For Phase 1, however, the process largely relied upon expert judgement and manual adjustment due to resource constraints and the current developmental status of the system.

TASK ANALYSIS

In parallel with use-case modeling, and in collaboration with Triton AVOs and other SMEs, a task analysis was conducted to better understand how AVOs respond to air-traffic encounters under Due Regard operations. The term 'hazard avoidance' describes this pilot behavior. Generally described, hazard avoidance includes two distinct but

overlapping goals: self-separation[3] and collision avoidance (SS/CA). The first goal is to maintain separation from other aircraft by providing an adequate buffer of time and space. This involves gradual and anticipatory AVO adjustments to the aircraft through new altitude, heading, or speed commands. The second goal, in the event that separation is lost, is to perform collision avoidance. In contrast to the deliberate and proactive nature of maintaining separation, collision avoidance is more reactive and undertaken with a sense of urgency.

Figure 2 shows a simplified task network with six tasks that comprise AVO hazard avoidance within an IMPRINT Pro project. Each task describes a discrete phase of the AVO response with distinct time and accuracy properties aligned to safety-case claim criteria. Hazard-avoidance behavior emerges as a result of the particular time and output values produced during a developing air-traffic encounter. Although depicted in a largely serial arrangement, an AVO, in practice, may perform the tasks non-sequentially and an overlapping fashion – in the interests of making the problem tractable, some concessions are made in this regard. An AVO begins hazard avoidance upon detection of air traffic, presented via an aircraft collision avoidance display (ACAD) [Task 1: Observe/Detect]. Note that due to other distractions brought about by the mission environment, the AVO may not immediately detect traffic displayed on the ACAD. After detecting an *intruder* (a term used to describe other aircraft), the AVO then performs an evaluation [Task 2: Evaluate Track]. In Task 2, the AVO carries out complex spatial reasoning, taking into account (among other factors) altitudes and airspeeds, and the range, relative bearing and heading of the intruder to make a prediction of whether there is a risk of losing safe separation or producing a collision. This process is dependent on the particular presentation of information on the ACAD. Next, if in an operational environment with multiple intruders, the AVO would perform a prioritization in an attempt to identify the most threatening intruder [Task 3: Prioritize Threats]. Following Task 3, the AVO decides on an action to resolve any threat identified during the evaluation task [Task 4: Decide on Action]. Actions range from "none," if there is no perceived threat, to deciding that a heading, altitude, or airspeed override is necessary to resolve the encounter. After having decided on a course of action, the AVO either commands the maneuver [Task 5: Command Maneuver] or, if no threat is perceived and no command required at the time, monitors the developing situation [Task 6: Monitor Encounter]. The AVO also typically monitors the encounter after having entered a new command in Task 5.

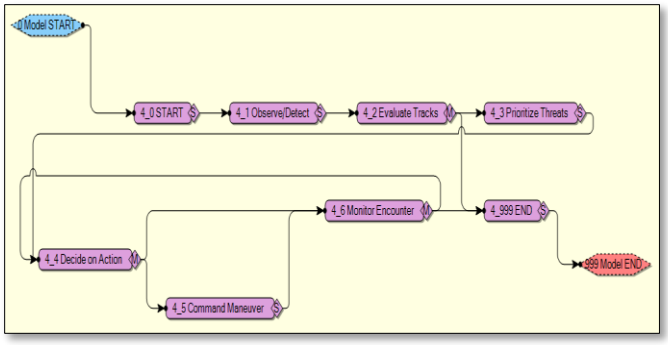

Figure 2: SAA Task Network in IMPRINT Pro

MODEL ARCHITECTURE

In accordance with NAVAIR-4.6 process standards, the OM architecture was developed from use-cases and related artifacts derived from the BAMS-D precursor, along with the associated task analysis and AVO subtree of the Triton safety case using object-oriented analysis and design (OOAD) techniques. Industry-standard design patterns were employed to promote portability, extensibility, and reuse. These patterns include the Visitor, Strategy, and Factory Method. Moreover, the design approach borrows from the Object Management Group's Model-Driven Architecture (MDA) methodology, reflecting the importance of domain engineering in this effort, as well as the need to remain technology agnostic to the extent practicable due to the wide range of tools and M&S environments employed on the program. Providing for future generalization to other aspects of platform operation was also a consideration.

Framework

Facilitating distributed and cross-platform implementations, the OM's framework was defined using a platform-independent model (PIM) with a high degree of decoupling among major components, achieved in part through use of the aforementioned design patterns. The framework is built around three major types of components: Task Model Units (TMUs), TAMs, and an SA class serving as the OM's working memory. Together, these elements may be

[3] Under Due Regard flight rules, an AVO is responsible for maintaining separation without air traffic control (ATC) by means of appropriate equipage. This is known as "self-separation."

composed to create a general executable task network supporting distinct, but interrelated facets of task performance. Additional detail is provided in the sections that follow.

Task Model Units

Embodying each AVO task identified during task analysis, the TMUs serve to aggregate a set of associated TAMs for execution within the context of a task network. As this association occurs at runtime, a TMU may be dynamically reconfigured with different TAM components, facilitating evaluation and comparison of alternative implementation strategies at the task level. TMUs also maintain a history of TAM results accumulated through multiple invocations of the simulation loop during task duration, enabling various schemes for combining intermediate results to produce final values upon expiration of the task timer. These values are passed out to the simulation as appropriate, used to update the model's SA, or may be logged for subsequent analysis. In addition, TMUs may apply a set of preconditions to the given world state and situation awareness, providing a guard against execution in an invalid situation.

Task Aspect Models

Providing the main functionality for task performance, the TAMs are divided into four categories, or aspects, aligned to the LVC task logs (see LVC Infrastructure section). As previously listed, these are duration, output, workload and successor. At the beginning of a task, the duration TAM provides an estimate (or prediction) for the interval during which the output and workload TAMs are invoked at a predetermined frequency (usually at the rate of the simulation loop) to provide intermediate results for processing by the associated TMU. The successor TAM then evaluates its path logic to determine the next branch to be taken in the task network. Task performance may be interrupted by critical changes in the world state or a discrete event. TAM functionality is implemented using a variety of strategies, including probabilistic models, heuristics, and machine learning. The choice of strategy depends on the nature of the task and aspect. In cases where the behavior may be described by a concise set of rules, a heuristic approach is often best. If the behavior is non-deterministic and largely independent of the current world state, a probabilistic approach may be used. For other situations involving complex patterns and many influencing factors, which is often the case for human tasks, ML models trained from LVC data may offer the only viable option. Figure 3 summarizes the TAM implementation strategies used for Phase 1.

	Task Aspect Models (TAMs)				
Tasks	Duration	Output	Workload	Successor	Comments
Detect/Observe	Independent PDF	Trivial SDT Heuristic (always 'Hit')	N/A for Phase 1	Output-based Path Logic	Convert to case-based PDFs on notification-type availability
Evaluate Tracks	Mixture Model w/ SOM-based Clusters	SME-based FIS	N/A for Phase 1	Output-based Path Logic	Tunable sensitivity; convert to ANFIS
Prioritize Threats	Independent PDF	Tau/DCPA Ranking Heuristic	N/A for Phase 1	Output-based Path Logic	Trivial execution for single-intruders
Decide on Action	Independent PDF	Hierarchical Probabilistic Classifiers	N/A for Phase 1	Output-based Path Logic	Probabilistic falloff oscillation damper for direction moderation
Command Maneuver	Case-based PDFs	Decision-coupled Heuristics, Case-based PDFs w/ Domain Constraints	N/A for Phase 1	Output-based Path Logic	Command coupled to decision w/ magnitude "defuzzification"
Monitor Encounter	Mixture Model w/ SOM-based Clusters	Multi-sample Probabilistic Classifier	N/A for Phase 1	Output-based Path Logic	ML training leverages multiple track samples separated in time

ANFIS – Adaptive Neuro-Fuzzy Inference System
FIS – Fuzzy-Logic Inference System
PDF – Probability Density Function
SOM – Self-Organizing Map

Figure 3: TAM Implementation Strategies

Situation Awareness

Acting as a perceptual buffer and working memory for the OM, the SA module stores pertinent information for task execution, which would be derived from the displays of the Mission Control System (MCS) console in actuality. Its design was informed through analysis of information exchanges documented by the use cases, together with feedback gathered during AVO interviews.

Constructive Integration

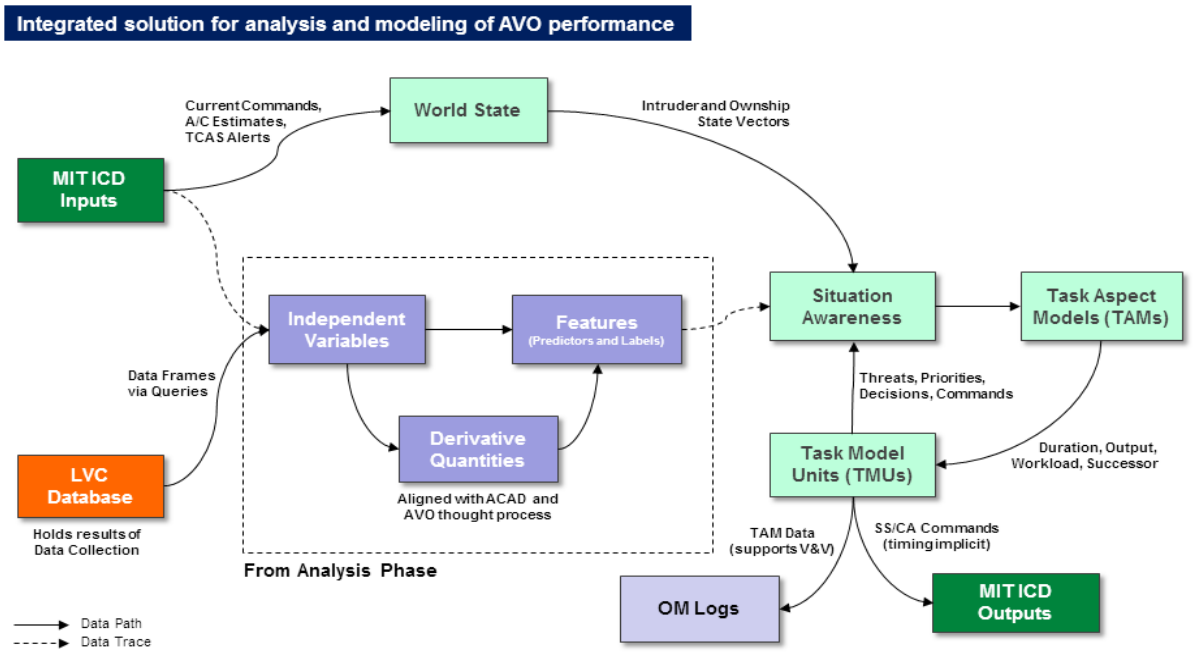

Figure 4: Operator Model Data Flows and CASSATT Integration

All inputs to the OM at runtime are specified by the Collision Avoidance System Safety Assessment Tool (CASSATT) interface control document (ICD), developed by the Massachussetts Institute of Technology (MIT) Lincoln Laboratory. Various transformation functions implemented within the model convert these inputs into world-state objects and SA data members for use by the TAMs. By design, these derived data match the semantics and representations of corresponding data-frame fields used to train the OM (see LVC Infrastructure section). The relationship between these data paths and model components is summarized in Figure 4.

Input values, intermediate quantities, and outputs from this process have been verified using the MIT test harness, test scenarios and logging features, and the OM V&V Test Tool (see Verification and Validation section). Collaborative integration testing with MIT has also supported data verification. Additionally, LVC data collected at the Triton Air Vehicle Environment (TAV-E) during trials have been verified through automatic checks, inspection of database tables, and comparison of visualization artifacts – so-called "spaghetti" diagrams generated by the OM V&V Test Tool, and ACAD screen captures received from the TAV-E. An example spaghetti diagram is shown in Figure 5.

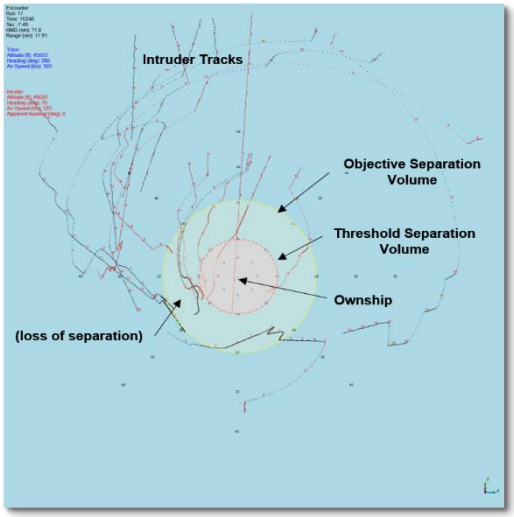

Figure 5: LVC Trials "Spaghetti" Diagram

LVC INFRASTRUCTURE

The following sections discuss the tools, processes, and procedures employed to conduct LVC trials, and collect the AVO and SAA performance data needed to support design, configuration, and training (in a machine-learning sense) of the OM. Their development and deployment comprised a significant portion of the total effort, which was undertaken with the intent of producing an enduring capability that could be leveraged, with appropriate customizations, to other UAS platforms.

AVO Performance Monitoring

The apparatus used to record an AVO's gaze vector, view, audio environment, and other associated data was the SMI Eye Tracking Glasses (ETG) sensor and mobile control unit. Following collection, the data were imported into the accompanying BeGaze™ software where it was integrated and reduced. Additionally, Intel RealSense™ equipment consisting of a 3D camera, microphone, and speech-recognition software recorded multi-aspect video, and converted AVO verbalizations to timestamped transcripts to support follow-on task coding (described below). Input and display loggers running on the control-station computers provided additional HCI data.

M&S Environment

The TAV-E, is a government-owned, high-fidelity M&S environment incorporating the Triton MCS Emulator and Closed Loop Simulator (CLS), both supplied by the Northrup Grumman Corporation (NGC). It provided both platform functionality and operationally-representative controls and displays sufficient to create a realistic and immersive experience of flight within the airspace. Prior to each run, AVO participants were given mission briefings and provided with information products representative of those normally generated during flight planning and mission development. Per the DOE, the information provided was controlled to prevent compromising the ecological validity of the experiment as an evaluation of SAA capabilities.

Data Supply Chain

Augmenting the TAV-E is the supporting infrastructure for the OM, which has been developed around a data supply chain, shown in Figure 6.

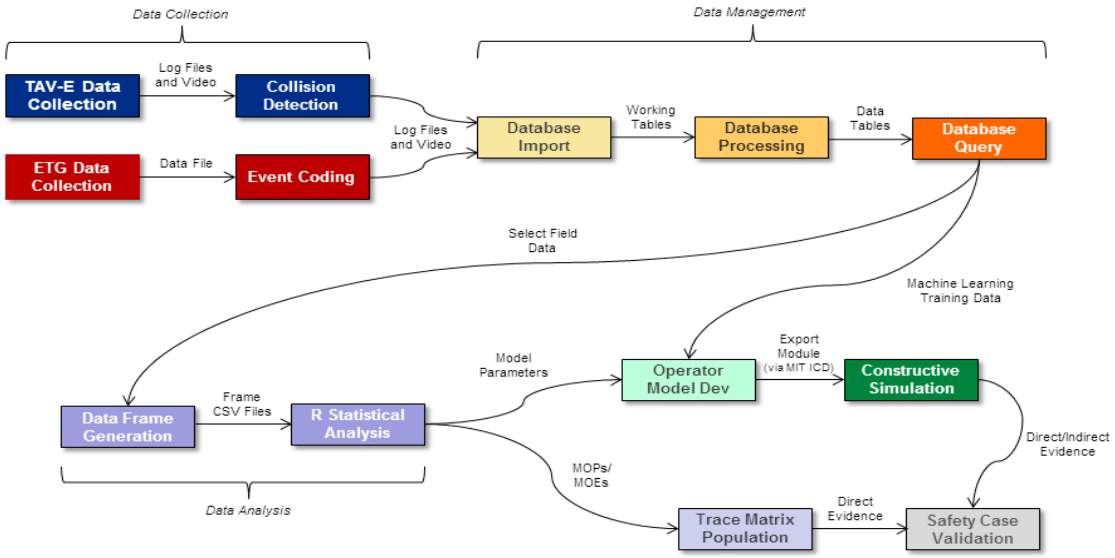

Figure 6: LVC Data Supply Chain

The chain begins with an LVC Data Collection segment, which monitors and records human response to HCI stimulus while operating the simulated MQ-4C Triton within the reference scenario. As described previously, it accomplishes this by means of various devices and software components, such as an eye tracker, digital video cameras, and a speech-recognition utility that collect performance data directly from the AVO participant. Moreover, through a semi-automated process referred to as *task coding*, an analyst interprets a subset of this data, including verbalized observations, evaluations, and intentions of the AVO regarding an intruder aircraft to infer task performance metrics traceable to the safety-case claims and MOPs specified in the AVO section of the Trace Matrix. The Data Collection segment also complements and leverages the virtual airspace environment provided by the TAV-E, including vendor-supplied Triton MCS Emulator and CLS software modules, as previously described.

Data reduction begins with the aforementioned task coding, in which an analyst reviews the ETG video and eye-tracking data collected from each LVC trial using the BeGaze™ software application, along with MCS screenshots and verbalization transcripts of observations, evaluations, and intentions from the AVO regarding intruder aircraft. From these data, the analyst segments the timeline into instances of specific task performance, and determines derived values for each field in a task log, using indicators and other guidance given by the Task Matrix – a product of the prerequisite task analysis. The Task Matrix, in turn, is aligned to the AVO section of the Trace Matrix, and specifically the MOPs specified therein. Fields comprising entries in the task log include duration, accuracy, output, workload, and successor type. The workload aspect is further divided into visual, auditory, cognitive, psychomotor, and speech (VACPS) subfields, in accordance with the Wickens' Multi-Resource Theory (MRT) (Wickens, 1984) and evaluation scale (Aldrich, 1989). Output syntax is unique to each task, and along with accuracy evaluation guidance, is specified in the Task Accuracy and Output Specification. Note that task duration and accuracy fields in particular trace to the safety claims criteria under the AVO subtree, while the task performance fields as a whole trace to the associated (indirect) safety-case evidence requirements. Given the complex set of artifacts and data elements involved, managing evidence tracing throughout the data collection and reduction processes is crucial. A set of custom UML-based diagrams has been devised and produced to trace relationships among data sources, database tables and queries, Task Matrix indicators, MOPs, and evidence requirements.

From here, log files containing scenario, system, and operator response data are passed to the Data Management segment, where they are imported, processed, and stored within a relational database. Following task coding, and the production of task logs for a simulated flight, the full set of text and multimedia logs obtained during data collection is imported into tables of a relational database, where they are timestamped in milliseconds of mission elapsed time (MET), before select field values are processed through conversion into the desired units, or transformation into the desired coordinate frames. Additional fields are then derived by update queries on the tables to facilitate later analysis. Another set of update queries then associates entries within the tables to realize table relationships specified by the database schema.

After this data management stage, the next phase of data reduction occurs, where a series of select queries are executed against the database to extract the designated fields that comprise the *data frame*. After extraction, the values are resampled into a synchronous format using a method appropriate to their type, which may be continuous, discrete, or interval. A final processing step then computes derivative quantities that capture the essential information presented on the ACAD and primary flight display (PFD), and intuitively align to cognitive task performance. The resultant data is now suitable for analysis in RStudio, where computation of summary statistics, distribution fitting, factor analysis, and feature extraction are performed.

The results of the Data Analysis and Management segments next inform the design, configuration, and training of the TAM components of the OM. As previously discussed, these components incorporate PDFs, heuristics, and ML elements such as self-organizing maps (SOMs), probabilistic neural networks (PNNs), and radial-basis function networks (RBFNs) to predict simulated task duration, output, workload, and successor values. Following implementation and testing, the completed OM is transitioned to MIT Lincoln for safety-case evidence generation via CASSATT.

VERIFICATION AND VALIDATION

A layered approach to model V&V was employed, including application of software best practices for testing at the unit and subsystem levels, use of the OM V&V Test Tool at the system level, and a validation strategy incorporating

qualitative assessments of response patterns, comparison of descriptive statistics for encounter outcomes, and statistical significance testing as described in the sections below.

OM V&V Test Tool

To support V&V of the OM at the NAWCAD, a custom software application was developed, referred to internally as the OM V&V Test Tool. Acting as a constructive-simulation surrogate for CASSATT, the test tool exercised the OM via execution against the same Phase 1 Run-Matrix encounters used for the LVC trials, thus providing a basis for comparison between the OM and AVO responses. Conforming to the CASSATT ICD, the test tool input dynamic world-state data to the OM (e.g., ownship and intruder position and velocity vectors), and received as output, AVO maneuver commands (e.g., heading and altitude overrides). After execution, typically in a stochastic batch mode, the tool logged simulation results using the same data-frame syntax as experimental phase, minus physiological response fields such as pupillometry. Together, the LVC and OM-based data frames supplied the validation strategy, discussed below.

Validation Strategy

The a priori validation plan involved comparing task performance data sampled from LVC trials with similar outputs from the OM using significance testing. However, some challenges exist with this validation strategy, stemming from the low number of observations in the LVC data set, mainly due to the dearth of available AVOs to participate in the experimentation. Compounding the issue was the requirement that the observation points be categorized by task and scenario prior to statistical testing to ensure only like observations were compared. However, it should be noted that this validation is an ongoing process - just as the model is being developed iteratively, the validation effort is also evolving.

Initial Validation Plan

The initial validation plan involved comparing task durations (the amount of time it took the operator to complete the tasks) and outputs (actions taken by the operator) using significance testing. Prior to significance testing, data were categorized first by task, then based on specific qualities of the scenario environment (i.e. ownship altitude, sensor, and sector of intruder approach), leading to categorization of data into 11 categories for each task.

The duration analyses compared the amount of time the AVO used to complete each task with the amount of time predicted by the OM. As is shown in Table 1, from an early version of the model, relatively few of the tests detected significant differences in duration at the alpha = .05 criterion level. For normally distributed data, Independent Samples *t*-Tests were used, whereas data that were not normal were tested using the Wilcoxon-Mann-Whitney U test.)

Duration Analyses by Task			
Subtask	Number of Tests Conducted	Number of Outcomes Significantly Different	Percentage of Outcomes Significantly Different
Detect & Observe	8	1	12.5%
Evaluate Tracks	10	2	20%
Decide on Action	9	0	0%
Command Maneuver	6	0	0%
Monitor Encounter	10	0	0%

Table 1: Overview of significance-testing results from duration analysis by subtask

In each task, the intention was to compare data based on the 11 scenario categories; however, these analyses were only conducted if there were a total of 10 observations (5 from the AVO, and 5 from the OM) within the category after outliers were removed.

For output analysis, output refers to the assessment, decision, or action the operator chose during a task, making the range of possible outputs dependent upon the task itself. For instance, in the Detect/Observe task, possible outputs included hit or miss, indicating whether or not the operator detected the intruder craft. Similarly, in the Evaluate Tracks

and Monitor Encounter tasks, the potential responses were dichotomous. However, in the cases of the Decide on Action and Command Maneuver tasks, there is a wide range of possible outputs, making meaningful analysis on this quantity of data unrealistic. For this reason, only the three subtasks with dichotomous outputs are evaluated herein. In the Detect/Observe task, the responses were constant for both the AVO and OM participants (hit). While this was expected since a missed intruder would be a very unlikely occurrence, it does indeed validate the model's behavior on this task. In the Evaluate Tracks task, the two possible outputs were "factor" and "no factor." The results in the Table 2 are reported in terms of percentage difference between the AVO and OM. As is shown the difference between the OM and AVO on this task is sizeable (note that these results are again from an early version of the model).

Difference in Responses on Evaluate Tracks Task											
Scenario Category	1	2	3	4	5	6	7	8	9	10	11
Percentage Difference	0.26	0.29	0.25	0.33	0.57	0.08	0.29	0.86	0.40	0.57	0.33

Table 2: Percentage difference between responses for the AVOs and OM on the evaluate tracks subtask

The Monitor Encounter task was evaluated in the same manner as the Evaluate Tracks task. The Monitor Encounter task yielded two possible outcomes: "threat condition" or "resume ops." Similar to the Evaluate Tracks task, many of the response differences here are sizeable (see Table 3).

Difference in Responses on Monitor Encounter Subtask											
Scenario Category	1	2	3	4	5	6	7	8	9	10	11
Percentage Difference	0.37	0.27	0.75	0.00	0.13	0.07	0.09	0.00	0.43	0.11	0.30

Table 3: Percentage difference between responses for the AVOs and OM on the monitor encounter subtask

While these differences in output indicated a need for finer tuning of the model and additional trials, they also suggested a need for improvements in the validation strategy. For this reason, updated validation analyses were planned.

Updated Validation Plan

To accommodate Phase 1 schedule constraints, the original validation strategy was modified to employ comparisons based on qualitative response patterns and descriptive loss-of-separation statistics for Objective, Threshold, and Standard volumes. In multiple, stochastic batch runs against the LVC Run-Matrix scenarios (29 encounters, with 30 iterations each), the delivered OM yielded an average separation loss rates of 38% Objective, 5% Threshold, and 0% Standard. The figures determined from the actual LVC trials were 42%, 5%, and 0%, respectively, with the caveat that the trial size was 65. No near mid-air collisions (NMACs) were observed from either source. Qualitative assessment of the output plots indicated a significant resemblance between the model and AVO responses, including features such as turn chains, turn directions, and turn reversals, as well as negative response to non-threat intruders.

CONCLUSIONS AND FUTURE DIRECTIONS

The effort was highly successful, providing a proof of concept and baseline capability for extension and enhancement in future increments. Moreover, valuable insights have been gained regarding actual AVO response in contrast to idealized representations derived from textbook procedures, which will help inform program decisions on training and HCI design. In particular, the observed variation in maneuver decisions with respect to altitude vs heading overrides, as well as blended maneuvers, suggest that significant reductions in separation loss rates could be achieved through emphasis on AVO training and use of suggestive displays employing recommended maneuver algorithms. Planning is currently underway to resume a more rigorous, task-based validation strategy, and revise the DOE to both expand the aperture of AVO participants, and collect a greater volume of data (with improved sampling) through part-task LVC simulations as appropriate. In addition, increased ecological validity for "black swan" events (Wickens, 2009), and investigation into the effects of workload manipulation, is expected as part of the progression in evidence-generation activities implementing the certification plan for the MQ-4C Triton platform.

REFERENCES

Abu-Mostafa, Y., Magdon-Ismail, M., Lin, H. (2012). Learning from Data. New York, NY: AMLBook.

Aldrich, T. B., Szabo, S. M. and Bierbaum, C. R. 1989, The development and application of models to predict operator workload during system design, in G. R. McMillan, D. Beevis, E. Salas, M. H. Strub, R. Sutton and L. Van Breda (eds), Applications of Human Performance Models to System Design, Defense Research Series, Vol. 2 (New York: Plenum), 65-80.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

ICAO, Official definition, ANNEX 6/I Operation of Aircraft — International Commercial Air Transport — Aeroplanes 2010.

Lutz R., Frederick P., Walsh P., Wasson K., Fenlason N. (2017), *Integration of Unmanned Aircraft Systems into Complex Airspace Environments,* Rapid Fielding of Capability to the Fleet, Volume 33, Number 4.

Nye, J. S. (2005), Soft Power: The Means to Success in World Politics, Perseus Book Group.

U.S. Navy Fact File (MQ-4C Triton), http://www.navy.mil/navydata/fact_display.asp?cid=4350&tid=500&ct=4, Updated February 15, 2017.

Wickens, C.D. (1984). "Processing resources in attention", in R. Parasuraman & D.R. Davies (Eds.), Varieties of attention, (pp. 63–102). New York: Academic Press.

Wickens, C.D., Hooey, B.L., Gore, B.F., Sebok, A., & Koenecke, C. (2009), Identifying black swans in NextGen: Predicting human performance in off-nominal conditions. *Human Factors*, *51*(5), 638-651.