Building a Human-Machine Teaming Training Testbed

Julia L. Walsh, Kent C. Halverson, Eric Watz
Aptima, Inc.
Woburn, MA
JWalsh@aptima.com, KHalverson@aptima.com, EWatz@aptima.com

David Malek
Wright State Research Institute
Dayton, OH
David.Malek@wright.edu

ABSTRACT

The Department of Defense faces increasingly complex mission sets, demanding unprecedented warfighter performance and flexibility in unpredictable situations. Autonomous intelligent agent (AIA) technology promises to significantly relieve these demands, in both training and operational settings. In the future, warfighters and AIAs will work jointly as part of mixed human-machine teaming (HMT) environments, wherein AIAs will routinely assess warfighter cognitive states to optimize learning, task management, feedback delivery, and performance. To meet the desired future state, training must equip humans with HMT skills necessary to optimize this collaboration. Learning environments must allow humans to experience authentic scenarios with HMT interactions that can be measured and assessed to provide real-time and after-action feedback.

To ensure HMT skill acquisition and retention, the AIA should meet the following system requirements. It needs to operate *autonomously* in complex environments, be *adaptable*, and be able to *enhance learning* by delivering strategic workload-based personalized feedback.

One way of achieving the aforementioned goals is by engineering a testbed that would meet the said system requirements and would train operators to interact with AIA and human teammates. The envisioned testbed would represent a fully deployable live, virtual, and constructive training system featuring integrated inner-loop (during the training scenario) and outer-loop feedback (after the scenario) capabilities along with a multi-modal measurement suite. The testbed would serve dual purpose – one for realizing operational training needs and the other for conducting within-subject multi-trial scenario-based studies to examine whether or not the presence of AIA impacts warfighter performance. Consistent with a growing body of literature which suggests that human performance tends to improve with an addition of an AIA (e.g., Mercado et al., 2016; McKendrick et al., 2014), we predict that warfighter performance will improve when AIA is present. While the research potential of the testbed is limitless, we offer several future research directions in this paper.

ABOUT THE AUTHORS

Dr. Julia L. Berger is an Associate Scientist in Aptima, Inc.'s Performance Assessment and Augmentation Division. Dr. Berger has theoretical knowledge and practical experience in developing and validating psychological instruments that measure attitudes, behaviors, and cognitions at various levels of analysis (e.g., individual, team, organization).

Dr. Kent C. Halverson is a Senior Scientist and Director of the Performance Assessment Technologies Division at Aptima, Inc. Dr. Halverson has research experience in multimodal performance measurement, organizational analysis, social network analysis, leadership development, and quantitative data analysis.

Eric Watz is Chief Engineer in Aptima, Inc.'s Product Engineering Group. Mr. Watz has over 16 years of professional experience in the development and integration of innovative software solutions for Department of Defense customers. As Chief Engineer, Mr. Watz oversees the technical strategy and vision for Aptima's core software products, including SPOTLITE®, PM Engine™, and A-Measure®.

Mr. David Malek is a Research Psychologist at Wright State Research Institute specializing in human factors, human performance, decision making, user-interface design, unmanned systems, and training. He currently leads WSRI's Live, Virtual, and Constructive (LVC) Simulation for training efforts aimed at first responders conducted at the National Center for Medical Readiness (NCMR) and other remote locations.

Building a Human-Machine Teaming Training Testbed

Julia L. Walsh, Kent Halverson, Eric Watz
Aptima, Inc.
Woburn, MA
JWalsh@aptima.com, KHalverson@aptima.com, EWatz@aptima.com

David Malek
Wright State Research Institute
Dayton, OH
David.Malek@wright.edu

INTRODUCTION

The Department of Defense faces increasingly complex mission sets, demanding unprecedented warfighter performance and flexibility in unpredictable situations. Autonomous intelligent agent (AIA) technology promises to significantly relieve these demands, in both training and operational settings. In the future, warfighters and AIAs will work jointly as part of mixed human-machine teaming (HMT) environments, wherein AIAs will routinely assess warfighter cognitive states to optimize learning, task management, feedback delivery, and performance. To meet the desired future state, training must equip humans with HMT skills necessary to optimize this collaboration. Learning environments must allow humans to experience authentic scenarios with HMT interactions that can be measured and assessed to provide real-time and after-action feedback. Thus, the current paper focuses primarily on discussing a concept testbed for training warfighters to interact efficiently and effectively with autonomous technologies in mixed HMT contexts.

In the present, the AIA technology is being heavily integrated in a wide range of military operations including casualty extraction, intelligence, reconnaissance, and surveillance (ISR), search and rescue (SAR), and others (Barnes & Evans, 2010; Greenemeier, 2010) with an assumption that the AIA augments human performance. However, burgeoning empirical evidence indicates that sometimes humans distrust autonomous agents; other times the AIA technology exerts undue cognitive demands on humans, suggesting that the AIA may potentially interfere with human higher-order cognitive functioning (Bitan & Meyer, 2007; Seppelt & Lee, 2007; Stanton, Young, & Walker, 2007). In fact, the literature describes the phenomenon known as the HMT paradox, in which humans and machines perform better independently than jointly (Trautman, 2017).

At the same time, AIA technology is becoming more sophisticated such that the human will soon become "the slowest element in the decision-making" process (Ryan, 2018). Therefore, it is imperative to understand how humans perceive and interact with AIAs to inform HMT training strategies and optimize learning and performance. Humans need to learn AIA behavior and cognition to be able to establish a mutually beneficial HMT relationship (de Visser, Cohen, Freedy, & Parasuraman, 2014, June; Lee & See, 2004). This is especially important in mission-critical environments where HMT ineffectiveness can lead to monetary loss or loss of human life.

To ensure HMT skill acquisition and retention, the AIA should meet the following system requirements. First, it needs to operate *autonomously* in complex environments. According to the Goal-Driven Autonomy (GDA) framework (Cox, 2013), which has generated interest in the research community as a means of allowing autonomous systems to operate for longer periods without human guidance, the agent should introspectively reason about own expectations and manage own goals in response to an evolving perspective on a dynamic, uncertain world (Aha, Klenk, Munoz-Avila, Ram, & Shapiro, 2010; Cox, 2013, 2015; Dannenhauer & Munoz-Avila, 2015). In other words, the agent should monitor its surroundings and recognize situational anomalies, which are often described as the difference between observation and expectation (Cox, 2013, 2015). Then, the agent should diagnose problems caused by the anomalies and generate goals to solve them (Klenk, Molineaux, & Aha, 2013; Paisner, Cox, Maynord, & Perlis, 2014).

Second, given the constantly changing nature of the mission-driven environments, the agents have to be *adaptable*. Most of the time, the AIA are designed with a certain set of parameters that can be challenged by novelty. The responses to these novelties, or 'surprises,' pose special interest to researchers and practitioners who try to balance computational robustness with optimization (Molineaux & Aha, 2014, July; Paisner et al., 2014, January).

Third, the AIA should *enhance learning* by delivering strategic workload-based personalized feedback (e.g.,

Adamczyk & Bailey, 2005, September). Humans are often faced with environments in which they have to multitask. Multitasking has been linked to an increased cognitive workload and error probability (Gontar, Schneider, Schmidt-Moll, Bollin, & Bengler, 2017). This is because the presence of multiple tasks leads to constant and unexpected interruptions. Therefore, if an AIA delivers real-time feedback, it is plausible that the interruptions caused by the agent may increase the human's cognitive workload such that he or she would need to switch attention from the current tasks to attend to the AIA feedback.

To avoid cognitive overload and reduce interruptions, the AIA should administer feedback at strategic time points and across appropriate modalities. There is a plethora of research investigating the effects of various interruption modalities (e.g., tactile, auditory, visual) and workload on HMT (e.g., Buettner, 2015; Lu et al., 2013). Some studies have shown that interruptions at low cognitive workload are known to be less disruptive than interruptions at high cognitive workload (measured by eye-tracking metrics; Katidioti, Borst, van Vugt, & Taatgen, 2016). There is further evidence to suggest that higher levels of AIA-support leads to lower user cognitive workload (Buettner, 2013, September). All this is to say that a major challenge in the HMT collaboration lies in the need to balance the AIA capabilities of delivering timely and relevant information with the human's ability to manage interruptions.

Finally, the AIA should be capable of participating in live (real participants operating real equipment), virtual (real participants operating simulated equipment), and constructive (simulated participants operating simulated equipment) training (LVC; Millar, Hodson, Peterson, & Ahner, 2016). Combined LVC training offers a number of benefits over live-only or virtual-only training. Among its most notable benefits are cost-effectiveness, on-the-fly scenario augmentation, high-fidelity, and logistical feasibility (Beaubien, Knapp, Wade, & Watz, 2017).

As mentioned before, the present paper focuses primarily on proposing a concept integrated simulation training testbed that would meet the aforementioned system requirements. Equipped with a multi-modal performance measurement (PM) suite, the envisioned testbed would represent a next-generation fully deployable LVC training system featuring integrated inner-loop (during the training scenario) and outer-loop feedback (after the scenario) capabilities (VanLehn, 2006). With these capabilities, the testbed would allow warfighters to train with AIA and human teammates and receive adaptive, personalized feedback during and after high-fidelity training scenarios.

In addition to bridging operational training gaps, the testbed would serve as a research platform to investigate empirical questions such as: How can AIAs effectively provide real-time, inner-loop, feedback via a novel user interface (UI) that provides personalized learning without overloading the human? What effect do various inner-loop feedback modalities (haptic, visual, auditory) have on task performance? While the research potential of the testbed is limitless, we offer several future research directions in this paper

CONCEPTUAL DESIGN OVERVIEW

As Figure 1 illustrates, the conceptual design for the testbed would need to include multiple components, the most important of which are: the Operator station; commercial off-the-shelf technology to output adversary and situational players as Computer Generated Forces; an AIA, which would computationally represent Compliance with Commander's Intent (CCI); the system-based PM capability; the observer-based PM capability; cognitive state assessment capability enabled by eye-tracking technology and data parsing algorithms; an AIA that would feature inner and outer loop feedback modules enhanced by a unique workload-based UI; chat-based communications capability with chat interpretation tool; and After Action Review (AAR) station with a replay capability.

It is worth mentioning that while the training station is geared for the Remotely Piloted Aircraft (RPA) use case, it would be largely role-agnostic, such that it could be repurposed for other roles, including an Air Warfare Coordinator (AWC) or a Tactical Action Officer (TAO) in the Combat Information Center (CIC) on a ship, and so on.

The system architecture leverages the principle of microservices, which employs small, focused software modules designed to ingest defined data structures. Each microservice within the architecture calls, processes, and stores its own data, and operates within its own process, thereby facilitating easy upgrades of the testbed components. Within a microservices architecture, the components may communicate using a synchronous protocol such as RESTful Web Services or an asynchronous protocol such as a Message Bus. To meet the data passing requirements, an open source message passing framework would need to be chosen that would allow for rapid message passing capabilities and includes a central capability to minor message traffic on multiple channels, or topics (i.e. ActiveMQ). Data within the

architecture are segmented into topics, with each topic having a well-defined JavaScript Object Notation (JSON) message format. Individual microservice components may subscribe to one or more topics within the testbed.

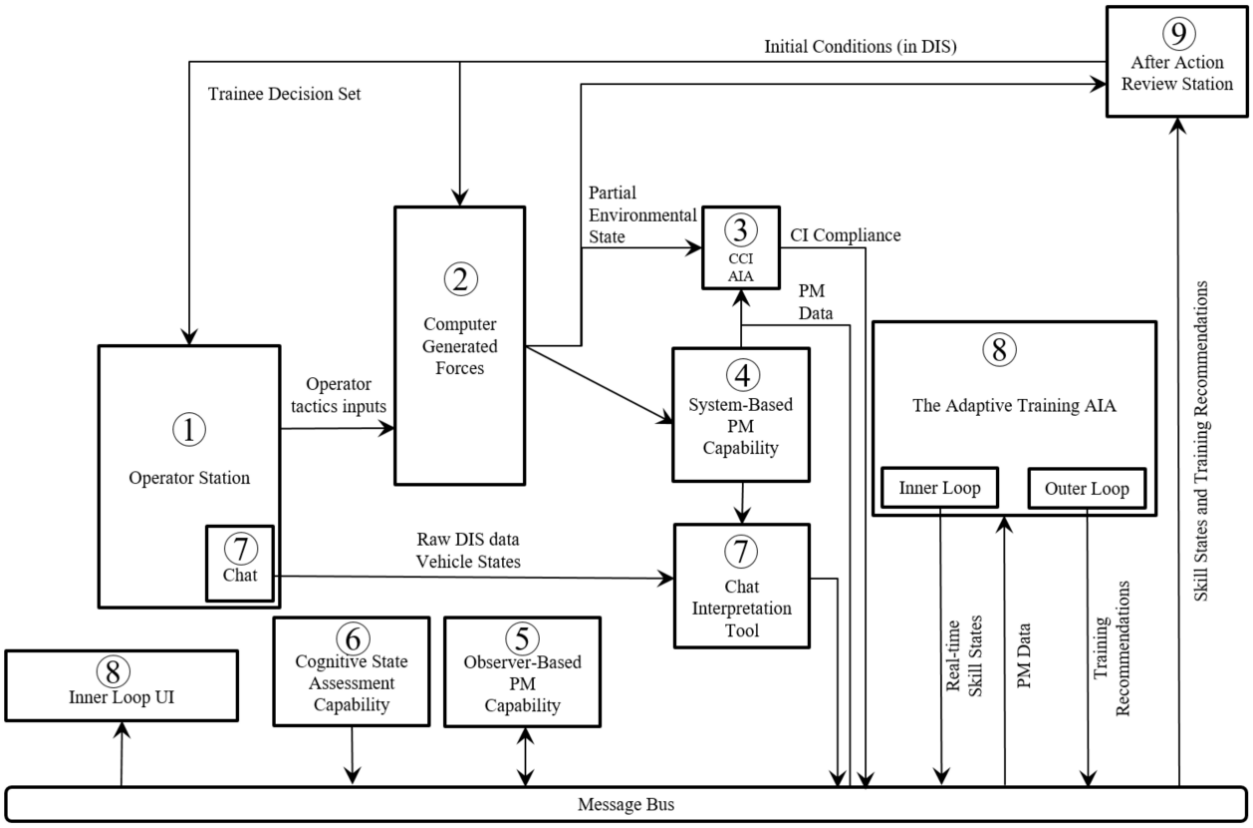

Figure 1. Conceptual Design

Below we detail the key components of the conceptual design that would uniquely position the testbed for practical and research opportunities in the area of personalized, adaptive HMT training.

CCI AIA

One of the key components of the testbed would be an AIA, which computes compliance with CI. CI is a statement that provides concise and effective guidance to subordinates of *what* is to be accomplished (as opposed to how) and *why* (Straight, 1995). As such, a formal CI data structure represents the highest level of guidance for learners in a training activity. Subsequently, learners interpret CI to determine appropriate course of action (tasking and priorities), develop goals, and execute behaviors consistent with the goals. Because the missions are often dynamic, CI may change and so do the goals. Therefore, the algorithms would need to ingest information about current CI, current environmental state, and the learners' choices and output a Compliance with CI (CCI) value between 0 and 1, where CCI closer to 1 would indicate consistency between the learners' choices and the CI.

System-Based Performance Measurement Capability

With an ever-increasing demand for the utilization of big data, especially in the applied settings, tools and technologies are being developed to capture human performance in an unobtrusive and objective ways, with the intent of optimizing warfighter performance. One example of these technologies is Air Force Research Laboratory's (AFRL) Performance Effectiveness/Evaluation Tracking System (PETS™), which is software that "enables multi-platform, multi-level measurement functionality at the individual and team levels in complex Distributed Interactive Simulation/High Level Architecture (DIS/HLA) environments" (Schreiber, Stock, & Bennett, 2006, p. 2). PETS is a successful Air Force Category 1 Advanced Technology Demonstration (ATD) project that has transitioned to operational use across multiple Air Force sites.

The primary driver for using a system-based PM tool like PETS in the envisioned testbed would be to capture human performance in real time and across different levels of analysis (individual and team). Outcome measures are mathematically-derived justifications of whether or not the mission objectives were met (Schreiber et al., 2006). Examples of PMs include but are not limited to task-appropriate aircraft speed, altitude, bearing, and time and duration of mission events. CCI AIA would use these and other PMs, in addition to environmental clues, to determine learners' CCI.

This capability would allow for the generation of automatic and objective data that "are difficult to argue with" (Schreiber et al., 2006, p. 1). Because objective measurement is devoid of human judgment, it describes individual actual performance (untainted by biases, prejudices, and preferences), helps diagnose the causes of success and failure, identifies gaps between actual and desired performance, and permits intra-individual and inter-individual comparisons cross-sectionally and longitudinally (Salas, Rosen, Held, & Weissmuller, 2009).

Observer-Based Performance Measurement Capability

Despite the undeniable advantages of objective PMs, an overreliance on objective measures may lead researchers to an incomplete understanding of the domain of interest. A latent attribute (construct), whether it be a warfighter's knowledge and/or skills (KS), can manifest itself in a number of ways. In addition, some constructs are only relevant under certain conditions created by a subtle and fleeting mix of numerous parameters. In this case, objective measurement may paint only a partial picture (explain some portion of variance) of the underlying constructs. Considering the complex, dynamic nature of some of the operational environments, algorithms for objective, system-based data is extremely difficult to develop and present to users.

Therefore, to complement the objective PM capability, the conceptual design for the testbed would need to include an observer-based PM capability. The capability would need to permit simulation instructors to capture and assess behaviors exhibited by individuals or teams in LVC training environments in real-time across a variety of performance measures (MacMillan, Entin, Morley, & Bennett Jr, 2013). For example, an observer would exploit this capability to report on how well the team followed CI using behaviorally-anchored ratings scales (BARS). Data obtained during a simulator session would allow analysis and presentation of performance data in debrief meetings following the exercise.

Adaptive Training AIA

The major goal behind the Adaptive Training AIA is to deliver adaptive, personalized trainee feedback during (inner-loop) and after training (outer-loop) in LVC training environments (Damian, Baur, & André, 2016, October).

The *inner loop* module would serve to deliver feedback to the learner in real-time to indicate a correct or incorrect response or provide an explanation of why a response is considered incorrect. At first, this module may serve to deliver performance feedback ("you failed on a given task"). Eventually, the inner loop process could also include an implicit recommendation to dynamically adapt the scenario to allow the learner to practice a failed task again.

As mentioned before, mission-critical environments require humans to process information in a short amount of time. This invariably increases the pull on their short-term memory retention capabilities, cognitive workload, and can affect decision-making skills. Any additional redundant information can lead to confusion and even errors or delays in executing tasks. To address this concern, an inner-loop feedback UI would need to be developed that would leverage recent findings in the interruption management and human cognition research. Specifically, it would rely on the findings of the recent research that indicates that integrating information from multiple sources into composite variables and effectively presenting them to the learners enhances human attention and cognition (Parasuraman, Sheridan, & Wickens, 2000; Wickens, Mavor, Parasuraman, & McGee, 1998; Wierwille & Eggemeier, 1998). The UI should minimize fatigue and cognitive overload, while providing the end users with the ability to query for knowledge in a manner consistent with that learner's function. Simply put, the cognitive agent (1) would consume the PM data from the observer-based and system-based capabilities to capture learning opportunities (low performance, errors); (2) would utilize the eye tracking technology to evaluate the learner's current cognitive workload (low, average, heavy); and (3) would deliver feedback to the learner about errors and opportunities to improve performance when workload is low.

The *outer loop* module would interpret learners' current skill states and recommends future training content to incrementally move the learners in the desired direction. For example, the adaptive training module could issue an AAR instructional policy to improve learners' competence related to CCI. This feedback would be delivered to the learners and instructor who may use these data for assessment, to drive classroom discussion, or to drive one-on-one feedback and lesson selection. To accomplish that, the adaptive training AIA would leverage models based on a domain-general mathematical approach and on the theory of deliberate practice (Ericsson, Krampe, & Tesch-Römer, 1993). It would contain adjustable parameters by looking across instances in which feedback has been applied and assessing the impact on learning. The underlying algorithms would be specially tuned to handle time-series data that are produced as the learners acquire and practice skills. The data produced by classic time-series algorithms depends only on a system state, which changes at each step. Research has extended these models to systems in which the data depend on decisions made by an intelligent agent (software or a human). In the adaptive training AIA, each state would be assigned a value and would use an algorithm, such as Reinforcement Learning, in real time to make decisions to maximize this value as the process unfolds.

Despite the fact that the above paragraphs describe the inner-loop and outer-loop capabilities as separate modules, their underlying algorithms perform the same function. Specifically, they estimate learner skill states and recommend learning content that would place the learners in the *zone of proximal development*, defined as the range of proficiency between what a learner can do independently and the upper limit of what they can accomplish with guidance and support (Vygotsky, 1987).

Taken together, the adaptive training AIA would interact with learners and other components of the testbed following the steps below:

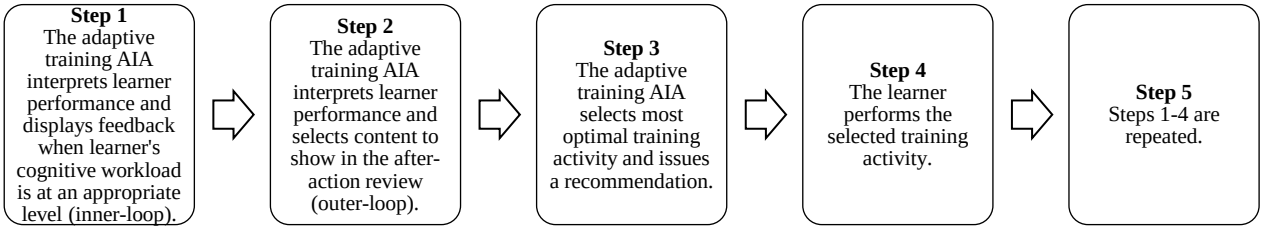

Figure 2. The Adaptive Training AIA Steps

Use Case (Scenario Development)

Before deploying the testbed in operational settings, all of its capabilities will need to be data validated on the basis of a use case. One example of such a use case may involve remotely piloted aircrews (RPA) that typically consist of a pilot, a sensor operator (SO), and a mission intelligence coordinator (MIC). Note that the recent transitions within the Air Force resulted in a partial reduction of the MICs, such that a good portion of the modern aircrews involves only a pilot and an SO. The current use case focused on the latter aircrew composition due to the seeming permanency of the aforementioned changes.

To demonstrate the two key capabilities included in the adaptive training conceptual design – inner-loop and outer-loop feedback – scientists may choose to engage the RPA aircrew in a search-and-rescue mission (SAR) within the contiguous United States (CONUS). Note that it is fairly atypical for the RPAs to respond to the like missions, with minor exceptions (e.g., military zones, lack of available civilian first responders). However, a key aspect of the domestic SAR scenario is the RPA aircrew acts as the on-scene commander, responsible for making tactical decision related to resource allocation, thereby providing copious learning opportunities. In fact, the number of decisions involved in a SAR mission and the fact that the RPA aircrew has the discretion to make those decisions, makes the SAR mission an appropriate scenario for the envisioned testbed.

The scenario development processes would be iterative in nature and would involve heavy collaboration with subject matter experts (SMEs). The scenario has to present a series of dilemmas (e.g., restricted airspace that creates time delays, bad weather that adds risk to the asset) to the aircrew that would require the aircrew to adjust its tactics.

Figure 3 illustrates hypothetical waypoints for the RPA. The scenario could begin with the aircrew flying an RPA, such as an MQ-9, from the initial RPA location to its default training location. While en route to the training site, the

aircrew would receive a mIRC chat (Internet Relay Chat) communication from Air Operations Center (AOC), which was notified by the Joint Personnel Recovery Center (JPRC) about a distress call from a lost hiker. The aircrew would be required to make a decision whether to respond to the call. Once the decision is made, the aircrew would be required to plan a route around restricted airspace as well as coordinate with other responding personnel and air assets (e.g., Pararescue team [PJs] on a helicopter). While en route to the target location, the aircrew would need to analyze the hiker location, terrain, and weather to select an orbit and the most appropriate search pattern. After locating the first hiker, the PJs would inform the aircrew of a second hiker moving east. At the same time, the aircrew would receive communication from the AOC via mIRC that the weather is getting worse. The aircrew would need to re-vector the RPA toward the suspected location of the second lost hiker. After finding the second hiker, the aircrew would communicate via mIRC with HH-60 helicopter. After that, the aircrew hands off control to the helicopter to initiate the rescue mission while returning to the base (RTB).

In Figure 3 below, a Google Earth screenshot showing hypothetical waypoints for the RPA is on the left. On the right is a MetaVR screenshot that shows a simulated RPA and the terrain. Note that the RPA aircrews' view would be different, such that they would view the terrain through their sensors.

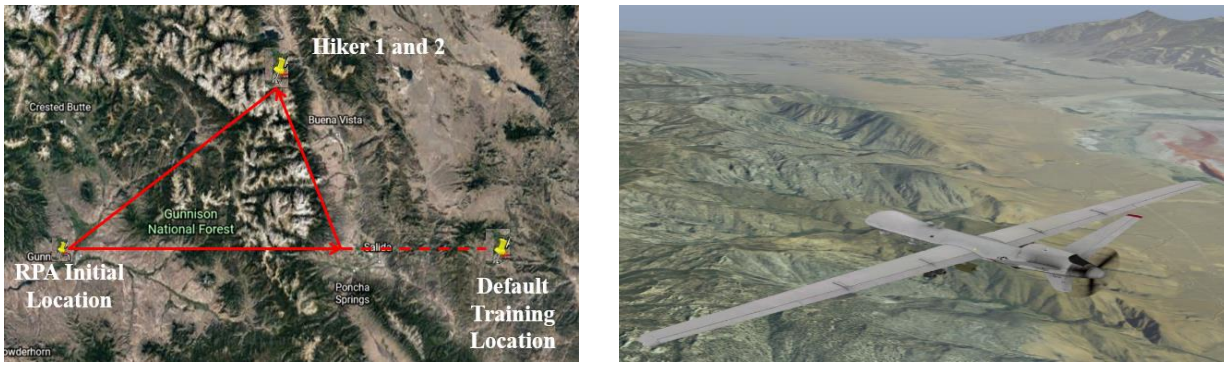

Figure 3. Google Earth Screenshot (left) and MetaVR Screenshot (right)

This hypothetical scenario includes several parameters that can be manipulated to generate variations of the rescue theme to support the development of a scenario library. The scenario library would entail variations of the one scenario such that each permutation can tailor attributes or parameters to focus on specific training objectives. The manipulated parameters may include the number and bearing of hikers, weather conditions, location of divert airports, and so on. All scenarios would represent real-world events while at the same time creating behavioral envelopes in which learners can be exposed to learning opportunities for exercising essential HMT skills. Thus, the adaptive training AIA would monitor scenario state and learner performance and provide feedback during scenario execution to ensure the learner experiences the intended training events.

Metadata Mappings

Optimal HMT performance is dependent on identifying the necessary competencies related to understanding and interacting with autonomous agents. HMTs are a relatively new concept and, therefore, lack well-defined knowledge/skills (KS) sets. To bridge this gap and to promote competency-based training, a thorough metadata mapping would need to be performed for the testbed to be fully functional. This mapping would serve as a necessary input to the adaptive learning algorithm intended to issue an instructional policy upon the completion of the scenario. The outcome of this mapping would be a linkage between PMs, KS, and tasks, such that performance on each task could be associated with a particular KS. To accomplish the mappings, several steps would need to be undertaken:

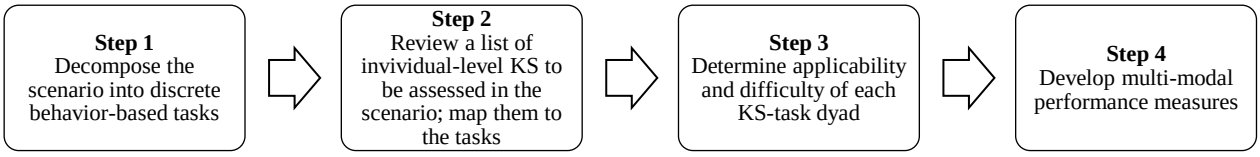

Figure 4. Steps Required to Accomplish Mappings

First, a list of tasks needs to be generated. Taking our hypothetical use case, the SAR scenario could be translated into 16 discrete behavior-based tasks.

Afterward, an exhaustive list of KS related to the RPA domain would be reviewed by the SMEs. To derive KS, it is recommended to utilize the artifacts of the Mission Essential Competences (MECs) workshops, defined as "higher-order individual, team, and inter-team competency that a fully prepared pilot, crew, flight, operator, or team requires for successful mission completion under adverse conditions and in a non-permissive environment" (as cited in Alliger, Beard, Bennett, Colegrove, & Garrity, 2007, p. 14). While the MECs involve team-level constructs, the KSs were not designed for team- or teams-of-teams-levels. In the absence of team-level KSs, individual-level KSs would need to be reviewed by the SMEs. Only those KSs that are related to team-level competencies would be kept. Each of the 16 behavior-based tasks would be associated with 3-4 KSs. Thus, of the available 98 KSs, a total of 19 KS would be associated with the 16 tasks. Note that the use case should guide the breadth and depth of the mappings.

Then, the SMEs would assist in determining the applicability and difficulty of each task-KS dyad using a 5-point Likert scale. For applicability, the anchors could range from 1 (very little applicability) to 5 (very high applicability). For difficulty, the anchors could range from 1 (very easy) to 5 (very difficult).

The adaptive training AIA would use these ratings for generating recommendations for future training content. In other words, the AIA would issue an instructional policy to the learners that would incrementally increase their proficiencies based on their skill state assessment from the preceding scenario. The instructional policy would be based on item response theory (IRT), which involves the strategy of presenting the learner with content that is sufficiently challenging, with a 50% probability of success. This strategy prevents under- or over-challenging the learner. Table 1 shows an example metadata mapping using the hypothetical use case.

Table 1. Sample of the Task-KS Matrix

TASKS		KS							
		Basic Piloting Skills		Compliance with Commander's Intent		mIRC Chat		Communication Standards	
		Appl	Diff	Appl	Diff	Appl	Diff	Appl	Diff
T1	Receive a briefing about the default mission from ATC			2	1	3	2		
T2	Receive a message about a lost hiker from JPRC			3	2	2	2	4	5
T3	Receive airspace clearance and permission from ATC to enter ROZ			4	3				

Note. T1-T3 – scenario tasks; KS – knowledge and skill; Appl. and Diff – applicability and difficulty. The table displays only 3 out of 16 tasks and 4 out of 19 KSs. Empty cells indicate that the given KS was not applicable. The KSs that were applicable for a given task were rated on a 5-point Likert applicability scale ranging from 1 (very little applicability) to 5 (very high applicability) and on a 5-point Likert difficulty scale ranging from 1 (very easy) to 5 (very difficult).

Development of PMs

Once the task-KS mapping is complete, a set of multi-modal PMs would need to be developed. It is recommended to utilize a standardized protocol that would be based upon a strong scientific and engineering foundation of time-tested PM methodologies such as the Rational Approach to Developing Systems-Based Measures (Orvis, DeCostanza, & Duchon, 2013), Competency-based Measures for Performance Assessment Systems (MacMillan et al., 2013), and Performance Effectiveness/Evaluation Tracking System (Schreiber et al., 2006).

Whatever protocol one decides to use, it is likely to be heavily SME-centric. Given our use case, the RPA domain is inherently complex and dynamic, involving the application of KS simultaneously, making it difficult to isolate and assess single aspects of performance. SMEs possess unique perspectives and understanding of the mission that would allow them to identify what aircrew success looks. After gaining insight from the SMEs, the researchers would

develop multiple PMs. The final list of PMs would be reviewed by the SMEs for content validity and by engineers for feasibility based on system requirements.

It is recommended to develop two groups of PMs – the first group would be achievable using existing software and simulation capabilities and the second group would involve future system capabilities, not yet available from the current testbed systems. The measurements that are not feasible to be implemented using today's tools or technology will become drivers for future system requirements. In addition, it is recommended to determine handshakes among the modalities and build the appropriate infrastructure where necessary because of the unique processes involved in each modality. Thus, for example, start conditions for some measures may be triggered by one modality and the end condition may be triggered by the third modality. This would require a careful engineering of the cross-modality handshakes.

As the result of the aforementioned steps, a metadata matrix similar to the one displayed in Table 2 could be created. The full matrix would contain information about each task, time from the beginning of the scenario until the task is terminated, KS number and description, and a detailed definition of each PM associated with each KS.

Table 2. Sample Metadata Matrix

T1 – Receive a briefing about the default mission from ATC Time – 0:01-1:00 KS1 – CCI										
PM#	Modality	Appl	Diff	SC	EC	PM	ARD	RS&R	SVC	Weight
PM1	system-based	2	1	As soon as the scenario starts	RPA receives a chat via mIRC from ATC	task appropriate aircraft speed	a continuous number from 0 to 480 km/hr	<= 300 is 0 (fail) 300 - 350 is 1 (pass) >= 350 is 0 (fail)	SV = RS/Range	.6
PM2	observer-based	2	1	As soon as the scenario starts	RPA receives a chat via mIRC from ATC	**How well did the aircrew follow CI?** 1- Did not follow CI 2- Followed CI incorrectly (prioritized the incorrect strategy) 3- Followed CI correctly (prioritized the correct strategy)	1, 2, or 3	1 - fail 2 - average 3 - pass	SV = RS/Range	.4

Note. T1 – Task 1; KS1 – Skill 1; CCI – Compliance with Commander's Intent; PM – performance measure; Appl. and Diff – applicability and difficulty; SC – start condition; EC – end condition; ARD - actual raw data (data acquired and transferred to the adaptive training AIA); RS&R - raw score & range (assessment of the learners' behavior in terms of passing or failing; SVC - standard variable created; Weight – weight assigned to each PM in the equation.

The RPA aircrew proficiency level (skill state) would be determined by a mathematical manipulation of the data received by the adaptive training AIA in the following order:

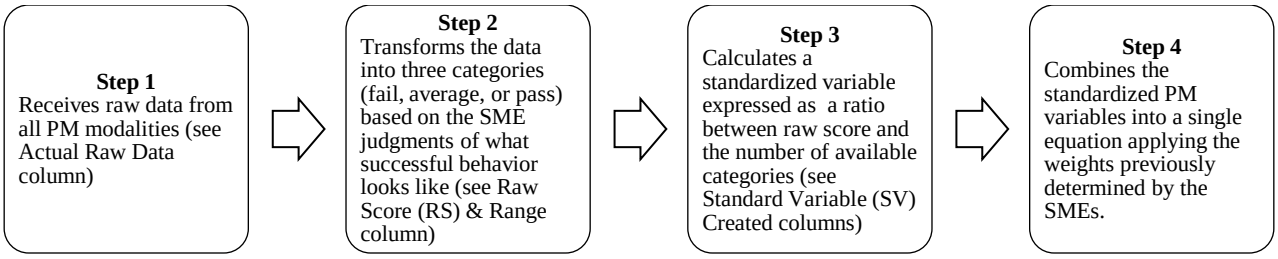

Figure 5. The Adaptive Training AIA Mathematical Data Manipulation

FUTURE RESEARCH DIRECTIONS

As mentioned before, the focus of the paper is on describing a concept LVC training testbed that would meet the desired future state where teams of humans and machines effectively and efficiently coordinate their activities. Equipped with a multi-modal PM metadata, inner-loop and outer-loop feedback capability, and algorithms to capture changing environmental states to assess CCI, the testbed would have both practical and research implications.

From the practical standpoint, the testbed would be unique in that it would leverage the AIA technologies that would be capable of operating autonomously for longer periods without human guidance and would be adaptable in the presence of novelties in the environment that violate core AIA expectations. Furthermore, the testbed would allow for high-fidelity LVC training for warfighters who are responsible for making high-stakes decisions in mission-critical mixed HMT environments.

From the research standpoint, once the testbed is built and fully operational, a research program can be commenced that would examine whether or not the presence of AIA impacts warfighter performance. Consistent with a growing body of literature which suggests that human performance tends to improve with an addition of an AIA (e.g., Mercado et al., 2016; McKendrick et al., 2014), we predict that warfighter performance will improve when AIA is present. We also predict that scenario complexity and cognitive workload will moderate this relationship. Furthermore, the testbed would be well positioned to investigate other empirical questions such as:

- General questions:
    a. How can system augmentations effectively provide real-time, inner-loop, feedback via a novel UI that provides personalized learning without overloading the human?
    b. Will information augmentation lead to accelerated KS acquisition?
    c. What types of input modalities and in what scenarios do they need to be presented to the user?
- Inner-loop-related questions:
    a. Examine the effect of inner-loop feedback modality (haptic, visual, auditory) and spatial separation on task performance or compliance with CI
    b. Is inner-loop feedback delivery (i.e., interruption) disruptive? If so, can the disruptiveness of interruption be mitigated with increased exposure to these interruptions?
    c. Which interruption management strategy is more effective – immediate, negotiated, mediated, or scheduled?
- Outer-loop-related questions:
    a. Does the adaptive training AIA impact participants' CCI?
    b. We can examine how the presence or absence of AAR impacts perceived training effectiveness
- Cognitive workload-related questions:
    a. How does workload change over time (within a trial/scenario)?
    b. Is there a noticeable change in workload following the receipt of a notification? If yes, is this change impacted by notification modality?

CONSIDERATIONS

As evident from the preceding sections, the envisioned testbed requires a well designed scientific and engineering (S&T) plan. Cost and logistics aside, there are a number of pitfalls to consider.

First, one needs to understand training requirements and gaps for a given domain to develop a robust use case. The requirements and use case will be used to inform an understanding of what training environment features must be present in order to produce the required conditions for learning. This competency-based approach increases the likelihood for incorporating the appropriate mix of technology within the overall training system design.

Second, one needs to determine the appropriate level of testbed fidelity from the start. Fidelity should be based on the extent to which the system enhances training quality rather than training realism (Stacy, Walwanis, Wiggins, & Bolton, 2013). This implies that training design should consider not only how well LVC entities resemble the real world objects, but also the outcomes the training tries to achieve. For example, simulated fire may not give off smoke in the same way as the real fire does, but if the goal of training is to master fire extinguishing skills, a laser simulator can be used which imitates fire behaviors in response to the laser-equipped fire extinguisher. We recommend using the layered fidelity framework (LFF) as a guiding principle for determining the elements of the blended-reality (BR) live, virtual, and constructed (LVC) training for Battlefield Airmen that would yield the highest learning utility and would enable fidelity assessment (Stacy et al., 2013).

Finally, to measure the full spectrum of human performance (i.e., to maximize the criterion space coverage), one would need to consider employing multiple data sources. This is because the full potential of each data source can be realized by triangulation. Triangulation is a process of selecting the most appropriate suite of modalities to gauge performance and later aggregating across these modalities to derive the well-rounded and data-informed conclusions. By doing so, the researchers and practitioners stand a better chance to control for the deficiencies inherent in each individual modality and to capitalize on their individual strengths. In theory, one may measure performance using all available modalities, but the reality of empirical investigations often places restrictions on what the practitioners can and cannot measure. Thus, scientists and engineers need to work closely together to consider resource constraints (i.e., funding, time) and environmental challenges (i.e., dynamic, multi-team, distributed) before settling on any given modality.

CONCLUSION

As mentioned before, the demand for effective and efficient collaboration between warfighters and AIAs is rapidly growing. The AIA technology promises to alleviate cognitive demands that warfighters often experience performing their duties in complex mission sets. To build the bridge between now and the future state wherein AIA work seamlessly with the humans in mixed HMT environments to augment human performance, we recommend engineering an LVC testbed that would serve both practical and research needs.

The testbed should be equipped with a number of AIA that operate *autonomously* in complex environments, be *adaptable*, and be able to *enhance learning* by delivering strategic workload-based personalized feedback. The testbed should also feature integrated inner-loop (during the training scenario) and outer-loop feedback (after the scenario) capabilities along with a multi-modal measurement suite. Combined, these capabilities enable the operators to learn to intact with AIA in mission-directed scenario-based environments. The adaptive learning capabilities would ensure continuous and proficiency-based learning at an individual- and team-levels.

ACKNOWLEDGEMENTS

The current effort was funded under contract #WSARC-16-00524 by Wright State Research Institute (WSRI) and the Air Force Research Laboratory (AFRL). The authors wish to acknowledge Mr. Bryan Davis, Mr. Logan Fuller, Mr. Benjamin Soltis, and Ms. Kate Heilner for their instrumental role as the SMEs in the initial phases of this capability development. We would like to acknowledge the work of the key WSRI researchers and engineers to include Dr. Michael Cox, Dr. Subhashini Ganapathy and their graduate students whose hard work and dedication to this project proved invaluable.

REFERENCES

Adamczyk, P. D., & Bailey, B. P. (2005, September). A method and system for intelligent interruption management.

In Proceedings 4th International Workshop on Task Models and Diagrams for User Interface Design, Gdansk, Poland.

Aha, D. W., Klenk, M., Munoz-Avila, H., Ram, A., & Shapiro, D. (Eds.) (2010). *Goal-directed autonomy: Notes from the AAAI workshop*. Menlo Park, CA: AAAI Press.

Alliger, G.M., Beard, R., Bennett, W., Colegrove, C.M. & Garrity, M. (2007). Understanding Mission Essential Competencies as a workload requirement. Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Readiness Research Division. AFRL-HE-AZ-TR-2007-0034.

Barnes, M. J., & Evans III, A. W. (2016). Soldier-robot teams in future battlefields: an overview. In *Human-robot interactions in future military operations* (pp. 29-50). CRC Press.

Beaubien, J.M., Knapp, M., Wade, A., & Watz, E. (2017). Performance measurement considerations for live, virtual, and constructive (LVC) training. In *Proceedings of the 2017 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. Paper No. 17091. Arlington, VA: National Training and Simulation Association.

Bitan, Y., & Meyer, J. (2007). Self-initiated and respondent actions in a simulated control task. *Ergonomics*, *50*(5), 763-788.

Buettner, R. (2013, September). Cognitive workload of humans using artificial intelligence systems: towards objective measurement applying eye-tracking technology. In *Annual Conference on Artificial Intelligence* (pp. 37-48). Springer, Berlin, Heidelberg.

Buettner, R. (2015). Investigation of the relationship between visual website complexity and users' mental workload: A NeuroIS perspective. In *Information systems and neuroscience* (pp. 123-128). Springer, Cham.

Cox, M.T. (2013). Goal-driven autonomy and question-based problem recognition. In *Second Annual Conference on Advances in Cognitive Systems 2013* (pp. 29-45). Palo Alto, CA: Cognitive Systems Foundation.

Cox, M.T. (2015). Toward a formal model of planning and action with goal reasoning. In *Goal reasoning: Papers from the ACS workshop* (pp. 37-51). GT-IRIM-CR-2015-001. AtlantaA: Georgia Tech.

Damian, I., Baur, T., & André, E. (2016, October). Measuring the impact of multimodal behavioural feedback loops on social interactions. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 201-208). ACM.

Dannenhauer, D., & Munoz-Avila, H. (2015, July). Raising Expectations in GDA Agents Acting in Dynamic Environments. In *IJCAI* (pp. 2241-2247).

de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014, June). A design methodology for trust cue calibration in cognitive agents. In *International Conference on Virtual, Augmented and Mixed Reality* (pp. 251-262). Springer, Cham.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, *100*(3), 363.

Gontar, P., Schneider, S. A. E., Schmidt-Moll, C., Bollin, C., & Bengler, K. (2017). Hate to interrupt you, but… analyzing turn-arounds from a cockpit perspective. *Cognition, Technology & Work*, *19*(4), 837-853.

Greenemeier, L. (2010). Machine self-awareness. *Scientific American*, *302*(6), 44-45.

Katidioti, I., Borst, J. P., van Vugt, M. K., & Taatgen, N. A. (2016). Interrupt me: External interruptions are less disruptive than self-interruptions. *Computers in Human Behavior*, *63*, 906-915.

Klenk, M., Molineaux, M., and Aha, D. (2013). Goal-Driven Autonomy for Responding to Unexpected Events in Strategy Simulations. *Computational Intelligence 29*(2): 187–206.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

Lu, S. A., Wickens, C. D., Prinet, J. C., Hutchins, S. D., Sarter, N., & Sebok, A. (2013). Supporting interruption management and multimodal interface design: Three meta-analyses of task performance as a function of interrupting task modality. Human Factors, 55(4), 697-724. doi:10.1177/0018720813476298

MacMillan, J., Entin, E. B., Morley, R., & Bennett Jr, W. (2013). Measuring team performance in complex and dynamic military environments: The SPOTLITE method. Military Psychology, 25(3), 266-279.

McKendrick, R., Ayaz, H., Olmstead, R., & Parasuraman, R. (2014). Enhancing dual-task performance with verbal and spatial working memory training: continuous monitoring of cerebral hemodynamics with NIRS. *Neuroimage*, *85*, 1014-1026.

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors*, *58*(3), 401-415.

Millar, J. R., Hodson, D. D., Peterson, G. L., & Ahner, D. K. (2016). Data quality challenges in distributed Live-Virtual-Constructive test environments. *Journal of Data and Information Quality (JDIQ)*, *7*(1-2), 2.

Molineaux, M., & Aha, D. W. (2014, July). Learning Unknown Event Models. In *AAAI* (pp. 395-401).

Orvis, K. L., DeCostanza, A. H., & Duchon, A. (2013). Developing systems-based performance measures: a rational approach. In The interservice/industry training, simulation and education conference (I/ITSEC, no. 1).

Paisner, M., Cox, M., Maynord, M., & Perlis, D. (2014, January). Goal-driven autonomy for cognitive systems. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Parasuraman, R., Sheridan, T.B., and Wickens, C.D. (2000). A model for types and levels of human interaction with automation. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 30*(3), 286-297.

Ryan, M. (2018, January). *Integrating Humans and Machines*. Retrieved on June 3rd, 2018 from https://thestrategybridge.org/the-bridge/2018/1/2/integrating-humans-and-machines

Salas, E., Rosen, M. A., Held, J. D., & Weissmuller, J. J. (2009). Performance measurement in simulation-based training: A review and best practices. *Simulation & Gaming*, *40*(3), 328-376.

Schreiber, B. T., Stock, W. A., & Bennett Jr, W. (2006). Distributed Mission Operations Within-Simulator Training Effectiveness Baseline Study. Volume 2. Metric Development and Objectively Quantifying the Degree of Learning. Lumir Research Institute Grayslake, IL.

Seppelt, B. D., & Lee, J. D. (2007). Making adaptive cruise control (ACC) limits visible. *International journal of human-computer studies*, *65*(3), 192-205.

Stacy, W., Walwanis, M., Wiggins, S., & Bolton, A. (2013). Layered fidelity: An approach to characterizing training environments. Proceedings of the 2013 Interservice/Industry Training, Education, and Simulation Conference, Orlando, FL.

Stanton, N. A., Young, M. S., & Walker, G. H. (2007). The psychology of driving automation: a discussion with Professor Don Norman. *International journal of vehicle design*, *45*(3), 289-306.

Straight, M. L. (1995). *Commander's Intent-An Aerospace Tool for Command and Control?*. NAVAL WAR COLL NEWPORT RI.

Trautman, P. (2017). A Mathematical Theory of Human Machine Teaming. *arXiv preprint arXiv:1705.03124*.

VanLehn, K. (2006) The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education. 16* (3), 227-265.

Vygotsky, L. (1987). Zone of proximal development. *Mind in society: The development of higher psychological processes*, *5291*, 157.

Wickens, C.D., Mavor, M., Parasuraman, R., and McGee, J. (1998). *The Future of Air Traffic Control* (Washington, DC: National Academy of Sciences).

Wierwille, W.W. and Eggemeier, F.T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35*(2), 263-281